THE FRONTIERS COLLECTION

Vesselin Petkov

RELATIVITY AND THE NATURE OF SPACETIME

Springer

THE FRONTIERS COLLECTION

THE FRONTIERS COLLECTION

The books in this collection are devoted to challenging and open problems at the forefront of modern physics and related disciplines, including philosophical debates. In contrast to typical research monographs, however, they strive to present their topics in a manner accessible also to scientifically literate non-specialists wishing to gain insight into the deeper implications and fascinating questions involved. Taken as a whole, the series reflects the need for a fundamental and interdisciplinary approach to modern science. It is intended to encourage scientists in all areas to ponder over important and perhaps controversial issues beyond their own speciality. Extending from quantum physics and relativity to entropy, time and consciousness – the Frontiers Collection will inspire readers to push back the frontiers of their own knowledge.

Information and Its Role in Nature
By J.G. Roederer

Relativity and the Nature of Spacetime
By V. Petkov

Quo Vadis Quantum Mechanics?
Edited by A. C. Elitzur, S. Dolev, N. Kolenda

Life – As a Matter of Fat
The Emerging Science of Lipidomics
By O.G. Mouritsen

Quantum–Classical Analogies
By D. Dragoman and M. Dragoman

Knowledge and the World
Challenges Beyond the Science Wars
Edited by M. Carrier, J. Roggenhofer, G. Küppers, P. Blanchard

Quantum–Classical Correspondence
By A.O. Bolivar

Mind, Matter and Quantum Mechanics
By H. Stapp

Quantum Mechanics and Gravity
By M. Sachs

V. Petkov

RELATIVITY AND THE NATURE OF SPACETIME

With 58 Figures

Springer

Dr. Vesselin Petkov
Concordia University
Liberal Arts College
de Maisonneuve Blvd. West 1455
Montreal, Quebec H3G 1M8, Canada
E-mail: vpetkov@alcor.concordia.ca

Cover figure: Detail from 'Venus Beauty and Anisotropic Geometric Diffusion' by U. Clarenz, U. Diewald, and M. Rumpf. Courtesy of M. Rumpf

ISSN 1612-3018
ISBN-10 3-540-23889-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-23889-8 Springer Berlin Heidelberg New York

Library of Congress Control Number: 2005924256

Springer is a part of Springer Science+Business Media.
springeronline.com

Typesetting by Stephen Lyle using a Springer TEX macro package
Final processing by LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover design by KünkelLopka, Werbeagentur GmbH, Heidelberg

Printed on acid-free paper SPIN: 10977378 57/3141/YL - 5 4 3 2 1 0

To all who struggle to understand this strange world

Preface

The standard books on relativity do not usually address the questions of the physical meaning of relativistic effects and the nature of spacetime. This book deals specifically with such conceptual questions. All kinematic consequences of special relativity are analyzed by explicitly asking whether the physical objects involved in these effects are three-dimensional or four-dimensional; this is equivalent to asking whether those objects exist only at the present moment of their times, as our common sense suggests, or at all moments of their histories. An answer to the question of the dimensionality of physical objects will resolve the issue of the nature of spacetime – whether spacetime is just a mathematical space (like a seven-dimensional color space, for instance) or represents a real four-dimensional world.

This book is intended for physicists, philosophers of science, philosophers, physics and philosophy students, and anyone who is interested in what special relativity is telling us about the world.

Acknowledgements

I would like to express my gratitude to all who contributed to the appearance of this book. So many people were involved in discussions on the issues covered here that it is virtually impossible to mention them all. That is why I would merely like to thank them for their constructive comments and hope that our discussions have brought us a little closer to understanding this beautiful but strange world.

I feel I should start the short list of specific acknowledgements by thanking Springer and Dr. Angela Lahee for starting the publication of The Frontiers Collection. I think the appearance of such a series is more than timely since scientists have already started to lose sight of new developments in the various scientific fields. I would also like to thank Stephen Lyle for his excellent technical editing of the manuscript.

I owe a lot to my teacher and friend Anastas Anastassov of Sofia University. His excellent lectures on general relativity in the 1980s

and our never-ending discussions prepared the ground for the ideas developed in this book. My thanks also go to Prof. Tzvetan Bonchev (at the time Dean of Sofia University's Faculty of Physics and Chair of the Department of Atomic Physics) and Prof. Ivanka Apostolova (at the time Chair of Sofia University's Department of Philosophy). Their influence is difficult to estimate.

I am grateful to my colleagues from the Department of Philosophy of Science of the Institute for Philosophical Research at the Bulgarian Academy of Sciences with whom many of the topics in this book were discussed in the late 1980s.

Versions of the issues examined in the book have been covered in different classes I taught – in the philosophy of science classes at Sofia University in the 1980s and later in the physics and in the philosophy of science classes at Concordia University. I am grateful to all students who participated in the class discussions. I also benefited from valuable comments from colleagues and students at Concordia University, McGill University and the University of Montreal, who attended a series of lectures I gave at a weekly seminar on General Relativity in the Fall of 1994, held at Concordia University. I would like to express my sincere thanks to all anonymous referees who made constructive recommendations and comments on different issues that are now included in this book. Most of the results presented here were also reported at several international conferences and at two inter-university seminars in Montreal – on open questions in physics and on the history and philosophy of science. I am truly grateful to the colleagues and students who took part in the discussions.

And last, I would like to express my deep gratitude to my wife Svetoslava and our son Vesselin (Jr) for their understanding, unconditional support, and encouragement. The completion of this book would not have been possible without their endless love and faith in me.

Montreal,
12 October 2004

Vesselin Petkov

Contents

Part I From Galileo to Minkowski

Part II On the Nature of Spacetime: Conceptual and Philosophical Issues

1 Introduction

This is not a typical book on relativity. It puts the emphasis on conceptual questions that lie beyond the scope of most physics books on this subject. The idea of such a book started to emerge more than twenty five years ago when I was struggling to *understand the meaning* of the consequences of special and general relativity. At that time I failed to find any physics books on relativity which addressed questions that looked so obvious to me. Here are three examples of such questions:

- It is stated in all books on special relativity that uniform motion is relative but no need has been seen to explain *why* absolute uniform motion does not exist. Answering this question is crucial for a genuine understanding of special relativity as the following apparent paradox demonstrates. Our common sense tells us that if a body moves *in* space it moves *with respect to* space. And indeed if we consider different examples of something moving in something else, it does appear that the expressions 'moving in' and 'moving with respect to' are equivalent. However, according to relativity such a conclusion is wrong since it is implicitly based on the idea of absolute motion. Therefore in relativity it is still correct to say that an object moves *in* space but not *with respect to* space. It is precisely here that the question of the non-existence of absolute uniform motion should be addressed in order to explain the profound depth of what lies behind the seemingly innocent difference between the two expressions.
- Another important issue that needs special attention is the physical *meaning* of the relativity of simultaneity. Logically, it comes after the question of absolute motion and can be approached differently depending on whether it is discussed in a physics or philosophy of physics class. In a physics class on relativity, my favourite problem for starting the analysis of what the physical meaning of the relativity of simultaneity is is the following:

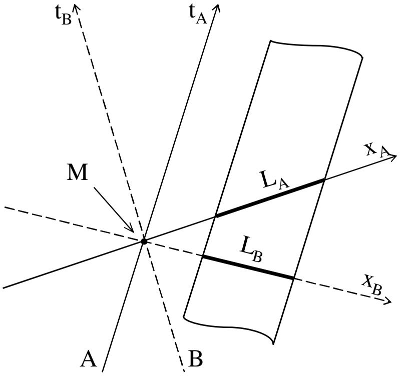
> An inertial reference frame S' moves with respect to another inertial reference frame S in the positive x direction of S. The clocks in S and S' are synchronized at the instant $t = t' = 0$ when the coordinate origins O and O' of the two frames coincide. At this moment a light wave is emitted from the point $O \equiv O'$. After time t it is observed in S that the light wave is spherical with a radius $r = ct$ and is described by the equation $r^2 = x^2 + y^2 + z^2$, which means that the center of the light sphere as determined in S is at O. Find the shape of the light wavefront in S' at time t'. Is it also a sphere whose center is at O'? If so, does this lead to a paradox? If not, does this lead to a contradiction with the principle of relativity?

The relativity principle requires all physical phenomena to look the same in all inertial reference frames. Therefore an observer in S' should determine that the wavefront of the propagating light signal is also a sphere whose center is at O'. This conclusion is confirmed by the Lorentz transformations. But our everyday experience tells us that there must be something totally wrong here – the center of the *same* light wave cannot be at two *different* places (at O and O' which may be thousands of kilometers apart). The standard explanation of this apparent paradox is the following: the wavefront of the propagating light sphere constitutes a set of *simultaneous* events and since according to relativity simultaneity is relative, the observers in S and S' have different sets of simultaneous events and consequently *different* light spheres. This is a correct explanation. But are you satisfied? I doubt it. This explanation is conceptually incomplete since it merely shifts the paradox from the specific case of light propagation to the relativity of simultaneity itself. What remains unexplained is *why* the two observers in S and S', who are in relative motion, have *different* sets of simultaneous events and therefore different light spheres (one centered at O and the other at O') given the fact that the two spheres originated from a *single* light signal. If the *physical meaning* of the relativity of simultaneity is explained conceptually then this apparent paradox will be explained as well.

- The above two questions as well as the question of the physical meaning of length contraction, time dilation, and the twin paradox all lead to the same major issue – how spacetime should be understood. Almost a century after Hermann Minkowski united space and time into an indivisible four-dimensional entity – now called Minkowski spacetime – the question "What is the nature of

> spacetime?" still remains open. In my view, this question should be addressed, not only in papers and books on the philosophy of spacetime, but in every physics book or university physics course on relativity. So far this has not been done, perhaps because most physicists seem to believe that their job is to make predictions which can be experimentally tested and that they need not bother about conceptual questions such as the following: Is Minkowski spacetime nothing more than a four-dimensional *mathematical* space which represents an evolving-in-time three-dimensional world or a mathematical model of a four-dimensional world with time entirely given as the fourth dimension? However, such conceptual questions cannot be avoided since the ultimate intellectual goal of all sciences, including physics, is to *understand* the world we live in.

In fact, even apart from pure intellectual curiosity, physicists themselves do need to address issues dealing with the interpretation of relativity if they want to offer some *explanation* of relativistic effects, which can make their mathematical description more transparent. Take for example length contraction as depicted in the figure below. Two inertial observers A and B in relative motion are represented by their worldlines (the lines of their entire lives in time). A meter stick is at rest in A's reference frame and is represented by its worldtube (its entire history in time) in the spacetime diagram shown in the figure.



The length of the meter stick is measured by A and B at event M when the observers meet, i.e., at the moment they set their clocks to zero: $t^A = t^B = 0$. As any length measurement requires that both ends of the meter stick be measured at the *same* time, and since A and B have different sets of simultaneous events, it follows that what A and B regard as their meter stick is, in fact, a *different* three-dimensional cross-section of the meter stick's worldtube. As the x axes of A and B intersect the worldtube at different angles, the two cross-sections L_A

and L_B are of different lengths, and this *explains* why A and B measure different lengths for the meter stick. The exact relation between the two lengths is obtained by the Lorentz transformations, which do show that $L_B < L_A$.

It is here that physicists cannot avoid the conceptual question of the nature of the meter stick's worldtube: Is the worldtube nothing more than just a graphical representation of the length contraction or a real four-dimensional object containing the whole history in time of the three-dimensional meter stick? It is clear from the spacetime diagram that, if we reject the reality of the worldtube of the meter stick, then A and B cannot have different cross-sections since only A's meter stick of length L_A would exist. This means that the same meter stick of the *same* length L_A would exist for B as well and no length contraction would be possible. Therefore the very existence of the relativistic length contraction seems to imply the reality of the meter stick's worldtube. This in turn implies the reality of Minkowski spacetime, since four-dimensional objects exist in a four-dimensional world.

Most books on relativity do not use spacetime diagrams specifically in the discussions of kinematic relativistic effects and do not face the immediate need to address the issue of the nature of Minkowski spacetime. Once obtained through the Lorentz transformations, these effects are not usually explained any further. In my view, such an approach is unsatisfactory for two reasons. Most importantly, physics is much more than its mathematical formalism and therefore everything should be done to provide a physical explanation of the results obtained through the Lorentz transformations. Secondly, if relativists themselves make no effort to shed some light on the meaning of the relativistic effects, different accounts start to emerge which in many cases are inconsistent with relativity itself.

One of the main reasons for writing this book is to address the issue of the physical meaning of the relativistic effects and the nature of spacetime by analyzing what the mathematical formalism of relativity is telling us. More specifically this is done:

- by carrying out an analysis of the idea of absolute motion starting from Aristotle's view on motion,
- by explicitly addressing the question of existence and dimensionality of the objects (rulers, clocks, twins, etc.) involved in the relativistic effects.

Part One entitled *From Galileo to Minkowski* starts with a chapter on the idea of absolute motion and how it was brought to its logical end by Galileo's refutation of Aristotle's view on motion. Chapter 3 is devoted to exploring the internal logic of Galileo's principle of relativity. I will argue that special relativity, and more precisely its four-dimensional formulation given by Minkowski, is *logically* contained in Galileo's principle of relativity (with a single additional assumption – that the speed of light is finite, which was determined experimentally in Galileo's century). An important result of this chapter will be the non-trivial conclusion that the non-existence of absolute uniform motion implies that the world is four-dimensional (or, equivalently, if the world were three-dimensional, absolute uniform motion had to exist because, as we will see in Chap. 3, a *single* three-dimensional world implies that 'moving *in* space' is equivalent to 'moving *with respect to* space'). Further exploration of the consequences of Galileo's relativity principle leads to all kinematic relativistic effects which are derived in Chap. 4. These derivations demonstrate that the relativistic effects are merely manifestations of the four-dimensionality of the world, whose geometry is pseudo-Euclidean, since these effects have direct analogs in the ordinary three-dimensional Euclidean space. One of the objectives of Part One is to show that special relativity could realistically have been formulated significantly earlier.

Part Two entitled *On the Nature of Spacetime – Conceptual and Philosophical Issues* is the most provocative of the three parts of the book. But it had to be written since the issues raised by the theory of relativity have challenged our entire world view in an unprecedented way. Never before has a scientific theory called for such a drastic revision of concepts that we have hitherto regarded as self-evident, such as the existence of:

- objective change,
- objective flow of time,
- free will.

In my view, special relativity has posed perhaps the greatest intellectual challenge humankind has ever faced. In this situation the best way to take on the challenge is to deal directly with its very core – the question of the nature of spacetime – since this question *logically precedes* the questions of change, flow of time, and free will. As we will see in Chap. 5, these issues crucially depend on what the dimensionality of the world is, which demonstrates that they are indeed preceded by the issue of the nature of spacetime.

For this reason the first chapter of Part Two (Chap. 5) examines the issue of the nature of Minkowski spacetime and argues that it is special relativity *alone* and the experimental evidence that confirms its predictions that can resolve this issue. This argument comes from the analysis carried out in the chapter which shows that special relativity is valid only in a four-dimensional world represented by Minkowski spacetime. Otherwise, if the world were three-dimensional, none of the kinematic relativistic effects would be possible, provided that the existence of the physical objects involved in the relativistic effects is assumed to be absolute (frame-independent). The only way to preserve the three-dimensionality of the world is to relativize existence. However, even this extreme step contradicts relativity itself and more specifically the twin paradox effect.

The profound implications of relativity (and its requirement that the world be four-dimensional) for a number of fundamental issues such as conventionality of simultaneity, temporal becoming, flow of time, free will, and even consciousness are also discussed in Chap. 5. It is shown that, in the four-dimensional Minkowski world:

- the definition of simultaneity is necessarily conventional,
- there are no objective becoming and time flow,
- there is no free will,
- the concept of consciousness (implicitly defined by Hermann Weyl [1] as an entity which makes us aware of ourselves and the world only at the moment 'now' of our proper time) is needed to reconcile the major consequence of special relativity that external reality is a timelessly existing four-dimensional world with the fact from our experience that we realize ourselves and the world only at the present moment.

It is these conclusions that constitute the intellectual challenge mentioned above. The most tempting way out of it is to declare them absurd or undoubtedly wrong. That is fine, if such a declaration is backed up by arguments demonstrating why those conclusions are wrong. A way to avoid facing the challenge is to subscribe to the view that we should accept the theory of relativity but should make no metaphysical pronouncement regarding the nature of spacetime. Such a view, however, completely ignores the fact that an analysis of the consequences of special relativity clearly shows that the challenge is there.

There exist two other approaches which try to avoid the challenge posed by special relativity. They purport to show that we should not bother about metaphysical conclusions drawn from special relativity for two reasons. According to the first approach the fact that relativity describes the world as four-dimensional and deterministic should not be taken as the whole truth since quantum mechanics, quantum gravity, and other modern physical theories are telling us different stories. Leaving aside the fact that quantum gravity and some of the modern physical theories are not yet accepted theories, Chap. 6 will make use of the results of Chap. 5 that it is the *experimental evidence* confirming the consequences of special relativity that contradicts the three-dimensionalist view. It would be really another story if the experimental evidence confirming the predictions of quantum mechanics contradicted the four-dimensionalist view. But this is not the case. Chap. 6 will present two arguments which demonstrate that quantum mechanics has nothing to say on the nature of spacetime.

Chapter 7 deals with the second approach according to which special relativity cannot tell us anything definite about the external world because, like any other theory, it may be disproved one day. We will see that this desperate attempt to avoid the challenge posed by relativity fails too. Again, this argument completely ignores the fact that it is the *experimental evidence* confirming the predictions of special relativity that contradicts the three-dimensionalist view. As experimental evidence cannot be disproved, any attack on the four-dimensionalist view should challenge the claim that experiment itself contradicts the accepted three-dimensionalist view. I will argue in this chapter that a scientific theory will never be disproved in its area of applicability where its predictions have been experimentally confirmed.

The main purpose of Part Two is to show convincingly that the challenge to our world view arising from special relativity – that the world is four-dimensional – is real. That is why it is only fair to face it now instead of leaving it for future generations.

Part Three entitled *Spacetime, Non-Inertial Reference Frames, and Inertia* further explores the consequences of the four-dimensionality of the world for physics itself. Chapter 8 starts by showing that relativity has resolved the debate over acceleration – whether it is absolute as Newton thought or relative as Leibnitz and Mach insisted. A body moving by inertia (with no acceleration) is represented in Minkowski spacetime by a straight worldtube; if the body accelerates, its worldtube is curved. Therefore, special relativity clearly shows that acceleration is absolute – there is an absolute difference between straight and curved worldtubes (and these worldtubes are, as argued in the book, not just convenient graphical representations but real four-dimensional objects).

The situation in general relativity is the same. The analog of a straight worldtube in a curved spacetime is a geodesic worldtube. A body moving by inertia (with *no curved spacetime acceleration*) is represented by a geodesic worldtube; if the body accelerates, its worldtube is deformed, i.e., deviated from its geodesic shape. Unlike relative velocity which cannot be discovered, an absolute acceleration should be detected experimentally. And indeed the propagation of light in a non-inertial reference frame, in which an accelerating body is at rest, turns out to be anisotropic – the average velocity of light depends on the body's acceleration. (The speed of light is c in all inertial reference frames in special relativity and in all local inertial reference frames in general relativity.) Most of Chap. 8 is devoted to the propagation of light in non-inertial reference frames – a topic that has received little attention up to now. The chapter ends with a discussion of the gravitational redshift effect and the Sagnac effect.

Chapter 9 shows that the potential and the electric field of a non-inertial charge can be calculated *directly* in the non-inertial reference frame in which the charge is at rest (without the need to transform the field from a comoving or local inertial frame) if the anisotropic velocity of light in that frame is taken into account. It is shown that the average anisotropic velocity of light in a non-inertial reference frame gives rise to a hitherto unnoticed anisotropic (Liénard–Wiechert-like) volume element which leads to the correct expressions for the potential and electric field of a charge in such a frame.

Chapter 10 addresses a natural question: If the deformed worldtube of an accelerating body is a real four-dimensional object, can the inertial force resisting the body's acceleration be regarded as originating from a four-dimensional stress in the body's worldtube which arises when the worldtube is deformed? It is argued in this chapter that inertia is another manifestation of the four-dimensionality of the world. Although the existence of inertia cannot be regarded as a definite proof of the reality of spacetime, it is shown in the chapter that, if the world is four-dimensional, inertia must exist.

2 On the Impossibility of Detecting Uniform Motion

One of the major events that marked the beginning of modern science in the seventeenth century was the acceptance of the heliocentric system of the world. In 1543 Copernicus [2] published his book on the heliocentric model of the solar system, but the acceptance of the new revolutionary view became possible only after the works of Kepler [3] and especially Galileo [4].

In this chapter we will see that Galileo played a crucial role in the Copernican revolution. He was the first scientist to apply systematically what we now call the hypothetico-deductive method (formulating hypotheses, deducing conclusions, and testing them experimentally) which is recognized as the key ingredient of a genuine scientific activity that leads to the formulation of a new theory. This approach helped him realize why Aristotle's view on motion had been the main reason for the dominance of the geocentric world system due to Aristotle and Ptolemy over the two preceding millennia. And indeed Aristotle's view on motion looked self-evident even in the seventeenth century since it appeared to be in perfect agreement with the common-sense view based on people's everyday experience. This view was almost certainly the ultimate reason for the rejection of the first heliocentric model put forward by Aristarchus of Samos (310–230 B.C.) immediately after Aristotle's geocentric system of the world.

With this in mind we can better appreciate Galileo's role in the acceptance of the heliocentric system. His disproof of Aristotle's view on motion was so important that one may wonder how many more years would have been needed for the ideas of Copernicus to be recognized if Galileo had not written his *Dialogue Concerning the Two Chief World Systems – Ptolemaic and Copernican.*

2.1 Aristotle's View on Motion

Aristotle did not hold any counter-intuitive views on motion as the Eleatics did.[1] His view reflected people's everyday experience and was summarized in the first sentence in Book VII of his *Physics*: "Everything that is in motion must be moved by something." Aristotle believed that there were two types of motion – natural motion of a body which tends to reach its natural place (the center of the Earth) and violent motion which is the motion that needs a mover. Aristotle himself realized that his view led to a problem since it could not explain the motion of projectiles [7, Book VIII, Chap. 10]:

> If everything that is in motion with the exception of things that move themselves is moved by something else, how is it that some things, e.g., things thrown, continue to be in motion when their movent is no longer in contact with them?

This is really an obvious argument against the way Aristotle explained motion: if we throw a stone it should stop at the moment it leaves our hand but this is not what is observed – the stone continues its motion *on its own* until it hits the ground. Aristotle seemed to believe that the observed continuing motion of projectiles can be explained by assuming that the medium in which projectiles travel is moving them. In the case of the stone it is our hand, while throwing the stone, that moves the medium (the air) which in turn acts as a mover of the stone.

Before discussing Galileo's crushing arguments against Aristotle's view on motion, let us examine in more detail how it contradicts the heliocentric system. Here is an excerpt from Ptolemy's *The Almagest* in which he employs Aristotle's view on motion in order to demonstrate that the Earth does not move [8]:

> Now some people, although they have nothing to oppose to these arguments, agree on something, as they think, more plausible. And it seems to them there is nothing against their supposing, for instance, the heavens immobile and the earth as turning on the same axis from west to east very nearly one revolution a day; or that they both should move to some extent, but only on the same axis as we said, and conformably to the overtaking of the one by the other.

[1] The Eleatic school of philosophy held that the observed motion and change are just illusions; the true reality, according to them, is an eternal existence [5, 6]. The Eleatic view is amazingly similar to the view suggested by special relativity, as we will see in Chap. 5.

> But it has escaped their notice that, indeed, as far as the appearances of the stars are concerned, nothing would perhaps keep things from being in accordance with this simpler conjecture, but that in the light of what happens around us in the air such a notion would seem altogether absurd. For in order for us to grant them what is unnatural in itself, that the lightest and subtlest bodies either do not move at all or no differently from those of contrary nature, while those less light and less subtle bodies in the air are clearly more rapid than all the more terrestrial ones; and to grant that the heaviest and most compact bodies have their proper swift and regular motion, while again these terrestrial bodies are certainly at times not easily moved by anything else – for us to grant these things, they would have to admit that the earth's turning is the swiftest of absolutely all the movements about it because of its making so great a revolution in a short time, so that all those things that were not at rest on the earth would seem to have a movement contrary to it, and never would a cloud be seen to move toward the east nor anything else that flew or was thrown into the air. For the earth would always outstrip them in its eastward motion, so that all other bodies would seem to be left behind and to move towards the west.
>
> For if they should say that the air is also carried around with the earth in the same direction and at the same speed, nonetheless the bodies contained in it would always seem to be outstripped by the movement of both. Or if they should be carried around as if one with the air, neither the one nor the other would appear as outstripping, or being outstripped by, the other. But these bodies would always remain in the same relative position and there would be no movement or change either in the case of flying bodies or projectiles. And yet we shall clearly see all such things taking place as if their slowness or swiftness did not follow at all from the earth's movement.

The above arguments can be summarized in a single argument discussed by Galileo in his *Dialogue Concerning the Two Chief World Systems – Ptolemaic and Copernican* published in 1632 [4, p. 139]. Consider dropping a stone from the top of a tower. If the Earth is not moving as the Ptolemaic view holds, the stone will fall at the base of the tower. Assume now that the Earth is moving (consider just its rotation). During the time a stone dropped from the tower falls the Earth will move and the stone will not fall at the base of the tower.

Since no one had ever observed such an effect the supporters of the Ptolemaic system maintained that the heliocentric system was wrong.

The arguments against the heliocentric system, which appeared to be so convincing for centuries, are based on Aristotle's view that everything that moves needs a mover. And indeed if we assume that the Earth is moving and we are on the top of the tower holding a stone, it does follow from Aristotle's view that the stone will stop moving with the tower at the moment our hand releases it – the mover (our hand) is not acting on the stone any more and it will stop moving in a horizontal direction. For this reason it will land at a given distance from the tower. At first sight such arguments appear irrefutable, and this is perhaps the most probable explanation for why the Ptolemaic system prevailed over the heliocentric system of Aristarchus of Samos.

2.2 Copernicus and Ptolemy's Arguments Against the Earth's Motion

In the sixteenth century Nicholas Copernicus (1473–1543) again argued that the Earth was not stationary at the center of the cosmos but rather rotated on its axis and also orbited the Sun. In his fundamental work *On the Revolutions of the Heavenly Spheres*, he advanced the argument that it was more natural to assume that the Earth is orbiting the Sun. However, as seen from the following quote he did not disprove Ptolemy's arguments against the Earth's motion [2, p. 519]:

> But let us leave to the philosophers of nature the dispute as to whether the world is finite or infinite, and let us hold as certain that the Earth is held together between its two poles and terminates in a spherical surface. Why therefore should we hesitate any longer to grant to it the movement which accords naturally with its form, rather than put the whole world in a commotion – the world whose limits we do not and cannot know? And why not admit that the appearance of daily revolution belongs to the heavens but the reality belongs to the Earth? And things are as when Aeneas said in Virgil: "We sail out of the harbor, and the land and the cities move away." As a matter of fact, when a ship floats on over a tranquil sea, all the things outside seem to the voyagers to be moving in a movement which is the image of their own, and they think on the contrary that they themselves and all the things with them are at rest. So it can easily happen in the case of the movement of the Earth that

> the whole world should be believed to be moving in a circle. Then what would we say about the clouds and the other things floating in the air or falling or rising up, except that not only the Earth and the watery element with which it is conjoined are moved in this way but also no small part of the air and whatever other things have a similar kinship with the Earth? Whether because the neighbouring air, which is mixed with earthly and watery matter, obeys the same nature as the Earth or because the movement of the air is an acquired one, in which it participates without resistance on account of the contiguity and perpetual rotation of the Earth.

Copernicus essentially *postulated* that all objects should participate in the Earth's motion. As the history of science has shown, this was not the best way to respond to an argument. Given the fact that Aristotle's view on motion was still the accepted doctrine in the sixteenth century, the arguments against the Earth's motion, which were based on Aristotle's view, were at that time valid arguments that had to be addressed properly. That is why the resurrection of the heliocentric system by Copernicus' ideas only became possible after Galileo disproved both Aristotle's view on motion and Ptolemy's arguments against the Earth's motion.

It is tempting to assume from this text that Copernicus implicitly advanced the idea of relative motion. A careful reading of his argument, however, shows that he simply wanted to point out that, just as it appears to the sailors that the harbor and the cities move away (whereas in fact it is the ship that is moving), it only looks to us that the heavens are rotating, whereas *in reality* it is the Earth that (absolutely) moves.

2.3 Galileo's Disproof of Aristotle's View on Motion

Galileo clearly realized that the arguments against the motion of the Earth and therefore against the heliocentric system were based on the Aristotelian doctrine of motion. For this reason he critically examined it and found it to contradict well-known facts about motion at that time. He did that in two independent ways. First, he showed that Aristotle's explanation of the motion of projectiles was wrong – in reality, once thrown, projectiles move on their own, not by the medium in which they travel. Second, he presented analyses of different experiments which independently arrived at the conclusion that in order to maintain their uniform motion, bodies do not need a constant mover. On the basis of the new view on motion, Galileo demonstrated that the arguments against the Earth's motion no longer hold, and this paved the way for the acceptance of the heliocentric model of the solar system.

Let us now see how Galileo achieved such an enormous result. In his *Dialogue Concerning the Two Chief World Systems – Ptolemaic and Copernican*, Simplicio defends the Ptolemaic system, whereas Salviati and Sagredo provide arguments against it.

First, Galileo gives an example of how a scientific debate should be conducted by stating the main arguments of his opponents. He does this through Salviati [4, p. 126]:

> As the strongest reason of all is adduced that of heavy bodies, which, falling down from on high, go by a straight and vertical line to the surface of the earth. This is considered an irrefutable argument for the earth being motionless. For if it made the diurnal rotation, a tower from whose top a rock was let fall, being carried by the whirling of the earth, would travel many hundreds of yards to the east in the time the rock would consume in its fall, and the rock ought to strike the earth that distance away from the base of the tower. This effect they support with another experiment, which is to drop a lead ball from the top of the mast of a boat at rest, noting the place where it hits, which is close to the foot of the mast; but if the same ball is dropped from the same place when the boat is moving, it will strike at that distance from the foot of the mast which the boat will have run during the time of fall of the lead, and for no other reason than that the natural movement of the ball when set free is in a straight line toward the center of the earth.

Now the stage is set for Galileo to show that these arguments against the Earth's motion are not irrefutable. As we will see the power of Galileo's arguments, presented by Salviati and Sagredo, is determined by the fact that they combine references to experiments and logical analysis. As one cannot perform the tower experiment on a moving Earth and on a motionless Earth to test whether it will produce different results, Salviati concentrates on the ship version of the experiment and asks Simplicio [4, p. 144]:

> You say, then, that since when the ship stands still the rock falls to the foot of the mast, and when the ship is in motion it falls apart from there, then, conversely, from the falling of

> the rock at the foot it is inferred that the ship stands still, and from its falling away it may be deduced that the ship is moving. And since what happens on the ship must likewise happen on the land, from the falling of the rock at the foot of the tower one necessarily infers the immobility of the terrestrial globe. Is that your argument?

After Simplicio agrees, Salviati continues [4, p. 144]:

> Now tell me: If the stone dropped from the top of the mast when the ship was sailing rapidly fell in exactly the same place on the ship to which it fell when the ship was standing still, what use could you make of this falling with regard to determining whether the vessel stood still or moved?

Simplicio's reply is: "Absolutely none". Salviati's next question is on whether Simplicio ever carried out "this experiment of the ship". He did not do it himself but insisted he believed the authorities "who adduce it had carefully observed it." At this point Salviati provides perhaps the clearest hint that Galileo performed the experiment with a stone falling from the mast of a moving ship [4, pp. 144–145]:

> For anyone who does will find that the experiment shows exactly the opposite of what is written; that is, it will show that the stone always falls in the same place on the ship, whether the ship is standing still or moving with any speed you please. Therefore the same cause holding good on the earth as on the ship, nothing can be inferred about earth's motion or rest from the stone falling always perpendicularly to the foot of the tower.

As Simplicio remains skeptical about what the result of a real experiment will be, Salviati virtually threatens him to make him realize the true conclusion without the need of any experiment [4, p. 145]:

> Without experiment, I am sure that the effect will happen as I tell you, because it must happen that way; and I might add that you yourself also know that it cannot happen otherwise, no matter how you may pretend not to know it – or give that impression. But I am so handy at picking people's brains that I shall make you confess this in spite of yourself.

What Salviati had in mind is the famous experiment involving inclined planes (see Fig. 2.1a) [4, p. 145]:

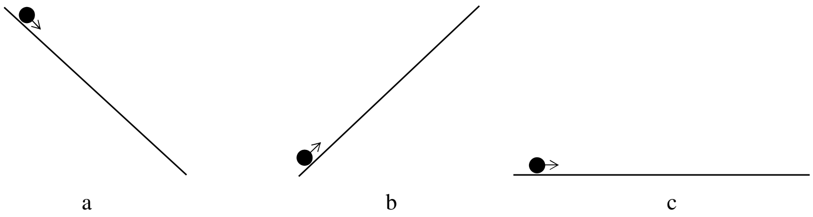


Fig. 2.1. Galileo's experiment with inclined planes

> Suppose you have a plane surface as smooth as a mirror and made of some hard material like steel. This is not parallel to the horizon, but somewhat inclined, and upon it you have placed a ball which is perfectly spherical and of some hard and heavy material like bronze. What do you believe this will do when released?

Simplicio gives the obvious answer: "the ball will continue to move indefinitely, as far as the slope of the surface is extended, and with a continually accelerated motion." Then Salviati asks what will happen to the ball if it is made to move upward on an inclined plane by a forcibly impressed impetus upon it (Fig. 2.1b). Simplicio does not have any difficulty responding to this question either [4, p. 146]:

> The motion would constantly slow down and be retarded, being contrary to nature, and would be of longer or shorter duration according to the greater or lesser impulse and the lesser or greater slope upward.

After discussing the two types of slope, Salviati takes the next logical step [4, p. 147]:

> Now tell me what would happen to the same movable body placed upon a surface with no slope upward or downward.

Simplicio seems to be a little perplexed [4, p. 147]:

> Here I must think a moment about my reply. There being no downward slope, there can be no natural tendency towards motion; and there being no upwards slope, there can be no resistance to being moved, so there would be an indifference between the propensity and the resistance to motion. Therefore it seems to me that it ought naturally to remain stable.

Now Salviati asks the crucial question [4, p. 147]:

> But what would happen if it were given an impetus in any direction?

Since Simplicio "cannot see any cause for acceleration or deceleration, there being no slope upward or downward," he unavoidably comes to the conclusion that the ball will continue to move "as far as the extension of the surface continued without rising or falling." This conclusion makes him agree with what Salviati said [4, p. 147]:

> Then if such a space were unbounded, the motion on it would likewise be boundless? That is, perpetual?

Salviati continues his argument [4, p. 148]:

> Now as that stone which is on top of the mast; does it not move, carried by the ship both of them going along the circumference of a circle about its center? And consequently is there not in it an ineradicable motion, all external impediments being removed? And is not this motion as fast as that of the ship?

After Simplicio admits that "this is true, but what next", Salviati urges him to [4, p. 148]:

> Go on and draw the final consequence by yourself, if by yourself you have known all the premisses.

Simplicio does see what follows from the premisses [4, p. 148]:

> By the final conclusion you mean that the stone, moving with unindelibly impressed motion, is not going to leave the ship but it will follow it, and finally will fall at the same place where it fell when the ship remained motionless.

However, he still refuses to accept the final conclusion and offers a counter-argument based on Aristotle's explanation of the motion of projectiles [4, pp. 149–150]:

> I believe you know that the projectile is carried by the medium, which in the present instance is the air. Therefore if that rock which was dropped from the top of the mast were to follow the motion of the ship, this effect would have to be attributed to the air, and not to the impressed force; but you assume that the air does not follow the motion of the ship, and is quiet. Furthermore, the person letting the stone fall does not need to fling it or give it any impetus with his arm, but has only

> to open his hand and let it go. So the rock cannot follow the motion of the boat either through any force impressed upon it by its thrower or by means of any assistance from the air, and therefore it will remain behind.

Simplicio fails to see the obvious – that the motion of the boat is impressed upon the stone by the hand of the person holding it; the stone is merely being pulled in the direction of the moving ship. As Simplicio's last defense is the issue of projectiles, Salviati has finally to deal with the weakest, but crucial element of Aristotle's view on motion – his account of what moves projectiles [4, p. 150]:

> Seeing that your objection is based entirely upon the non-existence of impressed force, then if I were to show that the medium plays no part in the continuation of motion in projectiles after they are separated from their throwers, would you allow impressed force to exist? Or would you merely move on to some other attack directed toward its destruction?

Simplicio agrees [4, p. 150]:

> If the action of the medium were removed, I do not see how recourse could be had to anything else than the property impressed by the motive force.

Before starting his attack on Aristotle's explanation of the motion of projectiles, Salviati asks Simplicio to state clearly Aristotle's view on "what the action of the medium is in maintaining the motion of the projectile" [4, p. 150], which he does [4, p. 151]:

> Whoever throws the stone has it in his hand; he moves his arm with speed and force; by its motion not only the rock but the surrounding air is moved; the rock, upon being deserted by the hand, finds itself in air which is already moving with impetus, and by that it is carried. For if the air did not act, the stone would fall from the thrower's hand to his feet.

Salviati then starts the formulation of his devastating argument [4, p. 151]:

> And you are so credulous as to let yourself be persuaded of this nonsense, when you have your own senses to refute it and to learn the truth? Look here: A big stone or a cannon ball would remain motionless on a table in the strongest wind, according to what you affirmed a little while ago. Now do you believe that

> if instead this had been a ball of cork or cotton, the wind would have moved it?

Not suspecting what will follow Simplicio confidently answers the question [4, p. 151]:

> I am quite sure the wind would have carried it away, and would have done this the faster, the lighter the material was. For we see this in clouds being borne with a speed equal to that of the wind which drives them.

Salviati asks Simplicio to answer one more question [4, p. 151]:

> But if with your arm you had to throw first a stone and then a wisp of cotton, which would move the faster and the farther?

Again Simplicio does not anticipate how much he is undermining his own position [4, p. 151]:

> The stone, by a good deal; the cotton will merely fall at my feet.

Now Salviati makes it impossible for anyone to defend what Aristotle assumed to be the cause for the motion of projectiles [4, p. 151]:

> Well, if that which moves the thrown thing after it leaves your hand is only the air moved by your arm, and if moving air pushes light material more easily than heavy, why doesn't the cotton projectile go farther and faster than the stone one? There must be something conserved in the stone . . .

This is one of most Galileo's brilliant arguments – he uses Aristotle's own explanation of how projectiles move to disprove this same explanation. It seems certain that Galileo recognized the crucial role of the issue of projectiles in Aristotle's view. In order that his arguments against Aristotle's explanation be as convincing as possible, he gave several arguments against it. Here is another devastating argument which this time is offered by Sagredo [4, p. 152]:

> But there is another point of Aristotle's which I should like to understand, and I beg Simplicio to oblige me with an answer. If two arrows were shot with the same bow, one in the usual way and one sideways – that is, putting the arrow lengthwise along the cord and shooting it that way – I should like to know which one would go the farther?

For one more time Simplicio is about to face the hidden contradictions between the Aristotelian doctrine of motion and our intuition obtained from everyday experience [4, p. 153]:

> I have never seen an arrow shot sideways, but I think it would not go even one-twentieth the distance of one shot point first.

Sagredo now exposes one of those contradictions [4, p. 153]:

> Since that is just what I thought, it gives me occasion to raise a question between Aristotle's dictum and experience. For as to experience, if I were to place two arrows upon that table when a strong wind was blowing, one in the direction of the wind and the other across it, the wind would quickly carry away the latter and leave the former. Now apparently the same ought to happen with two shots from a bow, if Aristotle's doctrine were true, because the one going sideways would be spurred on by a great quantity of air moved by the bowstring – as much as the whole length of the arrow – whereas the other arrow would receive the impulse from only as much air as there is in the tiny circle of its thickness. I cannot imagine the cause of such a disparity, and should like very much to know it.

Simplicio still does not seem to realize the contradiction [4, p. 153]:

> The cause is obvious to me; it is because the arrow shot point foremost has to penetrate only a small quantity of air, and the other has to cleave as much as its whole length.

Sagredo's explanation delivers the final blow to the view that it is the medium which continues to move projectiles after they are thrown [4, p. 153]:

> Oh, so when arrows are shot they have to penetrate the air? If the air goes with them, or rather if it is the very thing which conducts them, what penetration can there be? Do you not see that in such a manner the arrow would be moving faster than the air? Now what conferred this greater velocity upon the arrow? Do you mean to say that the air gives it a greater speed than its own?
>
> You know perfectly well, Simplicio, that this whole thing takes place just exactly opposite to what Aristotle says, and that it is as false that the medium confers motion upon the projectile as it is true that it is this alone which impedes it.

> Once you understand this, you will recognize without any difficulty that when the air really does move, it carries the arrow along with it much better sideways then point first, because there is lots of air driving it in the former case and little in the latter. But when shot from the bow, since the air stands still, the sidewise arrow strikes against much air and is much impeded, while the other easily overcomes the obstacles of the tiny amount of air that opposes it.

The conclusion that projectiles do not need a mover is inevitable. Once it becomes clear that projectiles move not by the medium but on their own, Aristotle's view – everything that is in motion must be moved by something – is essentially finished. The motion of an object which moves on its own is now called motion by inertia. It is controversial whether Galileo clearly realized the idea of inertial motion. Arguments which appear to show that he did not are easily found, mainly in the still rather Aristotelian terminology he used – motion "along the circumference of a circle about its center", "impressed motion", "impressed force", etc. What ultimately matters, however, is the essence of his arguments – that a body left on its own moves on its own and does not need a constant mover. And this is the very core of the fundamental idea of inertia. Galileo had tried to answer the question of why free bodies would continue to move on its own forever, provided that nothing prevents them from doing so, by assuming that the continued motion of a projectile is impressed upon it by its thrower. We have not done better than him – inertia still continues to be an outstanding puzzle in physics. The inertial motion of a body involves two questions:

- why does a free body move uniformly forever?
- why does a body resist the change in its uniform motion when it encounters an obstacle?

The first question will be addressed in Chap. 5, whilst Chap. 10 tries to outline a possible answer to the second.

2.4 Galileo's Principle of Relativity

Let us summarize the way Galileo disproved the arguments against the motion of the Earth. As these arguments were based on Aristotle's view on motion, Galileo carried out a brilliant analysis and convincingly demonstrated that, contrary to what Aristotle said, a body set free continues to move on its own without the need of a mover. Then

Galileo employed the new view of motion to both the tower and ship experiments and showed that a stone dropped from the tower or the mast of the ship preserves its motion and lands at the base of the tower or the mast, respectively. It is almost certain that Galileo carried out the experiment of releasing a stone from the top of a ship's mast and found that it always fell at the foot of the mast no matter whether the ship was moving or was standing still, which confirmed his arguments. In this way he demonstrated that experiments involving a stone dropped from a tower or from the mast of a moving ship always produce null results and therefore cannot be used to detect the motion of the Earth or the ship.

Therefore the motion of a body cannot be discovered by performing mechanical experiments (the type of experiments Galileo considered) on the moving body itself. Now we call this conclusion, which is derived from experimental facts, Galileo's principle of relativity: by performing mechanical experiments, the uniform motion of a body cannot be detected.

Before asking the question of the physical meaning of this principle in the next chapter let us end this chapter with another famous excerpt from Galileo's book which demonstrates the nullity of all experiments designed to show that the Earth is not moving [4, pp. 186–187]:

> For a final indication of the nullity of the experiments brought forth, this seems to me the place to show you a way to test them all very easily. Shut yourself up with some friend in the main cabin below decks on some large ship, and have with you there some flies, butterflies, and other small flying animals. Have a large bowl of water with some fish in it; hang up a bottle that empties drop by drop into a wide vessel beneath it. With the ship standing still, observe carefully how the little animals fly with equal speed to all sides of the cabin. The fish swim indifferently in all directions; the drops fall into the vessel beneath; and, in throwing something to your friend, you need throw it no more strongly in one direction than another, the distances being equal; jumping with your feet together, you pass equal spaces in every direction. When you have observed all these things carefully (though there is no doubt that when the ship is standing still everything must happen in this way), have the ship proceed with any speed you like, so long as the motion is uniform and not fluctuating this way and that. You will discover not the least change in all the effects named, nor could you tell from any of them whether the ship was moving or standing still.

In jumping, you will pass on the floor the same spaces as before, nor will you make larger jumps toward the stern than toward the prow even though the ship is moving quite rapidly, despite the fact that during the time that you are in the air the floor under you will be going in a direction opposite to your jump. In throwing something to your companion, you will need no more force to get it to him whether he is in the direction of the bow or the stern, with yourself situated opposite. The droplets will fall as before into the vessel beneath without dropping toward the stern, although while the drops are in the air the ship runs many spans. The fish in their water will swim toward the front of their bowl with no more effort than toward the back, and will go with equal ease to bait placed anywhere around the edges of the bowl. Finally the butterflies and flies will continue their flights indifferently toward every side, nor will it ever happen that they are concentrated toward the stern, as if tired out from keeping up with the course of the ship, from which they will have been separated during long intervals by keeping themselves in the air. And if smoke is made by burning some incense, it will be seen going up in the form of a little cloud, remaining still and moving no more toward one side than the other. The cause of all these correspondences of effects is the fact that the ship's motion is common to all the things contained in it, and to the air also.

3 Exploring the Internal Logic of Galileo's Principle of Relativity

To the memory of Sava Petrov[1]

> The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.
>
> H. Minkowski [9, p. 75]

This chapter pursues two aims. First, to address the important question of the physical meaning of Galileo's principle of relativity. Second, to demonstrate the power of an effective method of scientific enquiry – exploring the internal logic of fundamental ideas. In the previous chapter we have seen an excellent example of how this method works and how it produces results with far-reaching implications. Here we will follow Galileo's approach by studying his principle of relativity and more specifically its physical meaning. As we will see, the exploration of the internal logic of Galileo's principle of relativity will reveal its profound depth and will help us to arrive at the quite unexpected result that the theory of relativity (more specifically its four-dimensional formulation

[1] In the 1980s Sava Petrov, Head of the Department of Philosophy of Science in the Institute for Philosophical Research at the Bulgarian Academy of Sciences, proposed an ambitious research project – exploring the internal logic of scientific ideas. The goal was to provide philosophers of science with a powerful method that could ideally allow them to outline directions of research in the various sciences and not only interpret the already existing scientific discoveries. At that time the main results of this chapter were presented as an illustration of how this method might work.

given by Hermann Minkowski) is *logically* contained in Galileo's relativity principle. Such a result implies that special relativity could have been discovered earlier. Here we will not speculate on when it could have happened. We will simply carry out the analysis of Galileo's principle on the basis of what we now know and show that it does contain special relativity in its four-dimensional formulation in the sense that it can be deduced from Galileo's conclusion that the uniform motion of a body in space cannot be detected with the type of experiments Galileo considered. Once this has been done, one can indeed ask the question: When could the theory of relativity realistically have been discovered?

3.1 On the Physical Meaning of Galileo's Principle of Relativity

By disproving Aristotle's view on motion and providing irrefutable arguments supported by experiments, Galileo succeeded in demonstrating that any experiments of the type he discussed designed to show that the Earth is not moving would always produce null results. In other words, no matter what kind of mechanical experiments we perform, we cannot detect the uniform motion of a body in space. Let me specifically emphasize that Galileo did not disprove the Ptolemaic system of the world; what he disproved were the arguments against the Copernican system. The latter was accepted as the correct description of the world on the basis of a combination of reasons, mostly astronomical observations, but also the requirement of logical simplicity – a single assumption that it is the Earth that orbits the Sun explains all astronomical observations without the need to introduce epicycles or other ad hoc hypotheses.

Galileo's principle of relativity simply states the null results of the experiments intended to detect the uniform motion of an object, but does not answer the fundamental question of *why* we cannot discover that motion. The way this principle is stated – by performing mechanical experiments we cannot detect our uniform motion in space – seems to imply that there is such a motion but we cannot discover it. At the very moment we start to ask this type of question – which marks the beginning of what we called exploring the internal logic of Galileo's relativity principle – we are overwhelmed by an avalanche of other questions and apparent paradoxes.

The first profound question is: If the Earth is moving, what does it move in? Our everyday experience tells us that objects move in different media – in air, water, etc. Therefore it is natural to assume that the Earth is also moving in some kind of medium, which can be called aether, vacuum, or simply space. Now we face the first problem: Is space an entity? Since the time of Newton and Leibnitz there has been a continued debate over the nature of space. There exist two opposing views – absolutism (substantivalism) and relationism, which hold that space is a substance and a collection of relations between physical objects, respectively. In this book we will not enter that debate but, in order to avoid semantic misunderstandings, we will make it clear that by space we will understand the entity in which the Earth moves. One cannot deny the existence of that entity by claiming that it is merely nothingness or just a set of relations between the Earth and the other celestial bodies. Such a denial must obviously answer the old argument – if there were nothing between the Earth and the Moon, for example, why would they not be touching each other? The relationists should also explain how the Earth can move in a collection of relations between objects.

If space is an entity, it is obviously absolute in the sense that, being *one* entity, space is for everyone just *the* space. Then, any motion in space is an absolute motion; any rest in space is absolute rest. It was in this sense that motion and rest had been regarded as principal conditions in nature from the time of Aristotle to the time of Galileo [4, p. 130]. Now we face another problem – if space is regarded as a medium, as in any other known medium, the natural state of objects in space should be to be at rest, and any moving object should require a mover since media impede the motion of objects. This problem helps us to better understand and appreciate Aristotle's view. What is wrong then? What caused that problem? One may be tempted to postulate that space is a different kind of medium which does not resist the motion of objects, and for this reason no mover will be necessary. The obvious question arising from such an ad hoc hypothesis is how a medium (one of whose basic features is to impede the motion of any object) could offer no resistance to moving objects.

In order to see that the above ad hoc hypothesis cannot solve the problems which an absolute space causes, let us continue our examination of Galileo's relativity principle. Imagine that in the period between the time of Galileo and twentieth century there existed a group of scientists as bright as Galileo himself. Those scientists had been studying Galileo's arguments and wanted to develop them further. A natural beginning would be to try to understand the physical meaning of his relativity principle by asking why it is impossible to discover the uniform[2] motion of the Earth in space, if it is moving, or the uniform motion of a ship (everyone sees that the ship is moving in space). After some discussion they would arrive at the only two logically[3] possible interpretations:[4]

- There is absolute uniform motion but we cannot detect it.
- There is no absolute uniform motion and that is why we cannot detect it.

At first sight it appears that the first interpretation of Galileo's principle of relativity is almost certainly the correct explanation of the physical meaning of his relativity principle. However, upon closer examination it becomes evident that it leads to a lot of problems, including the one above – that space should offer no resistance to moving objects. On the other hand, however, the second interpretation seems even worse as we will see below. That is why our scientists decide to split into two research teams, each studying one of the interpretations of Galileo's relativity principle. The reader can follow the analyses of the members of the first more traditional research team who study the first interpretation. Here we will for the main part be carrying out the analyses that the second more radical research team had to perform. There are two reasons for this choice. First, we know that the second interpretation of Galileo's relativity principle turned out to be the correct one. Second, we also know from the history of science (which we should take into account especially in view of what Hegel warned – that we learn from history that no one learns from history) that any time there are competing explanations of a difficult problem, the most radical one has the best chances of being the true explanation. When looking at the two interpretations of Galileo's principle of relativity, one may really wonder what could be more radical than declaring that there is no absolute uniform motion. Let us see why.

[2] For a short period of time (for example, the time necessary for a stone dropped from a tower to reach the ground) the motion of the Earth is a good approximation to uniform motion.

[3] It seems that a third interpretation, viz., that there is absolute motion but that it cannot be discovered with mechanical experiments, is possible, but this interpretation will be examined later and it will be shown that, if absolute uniform motion existed, it should be discovered with mechanical experiments as well.

[4] What we have been trying to do here is to reconstruct a hypothetical analysis that could have been carried out earlier than the twentieth century. Although it is extremely difficult to eliminate any help from what we now know, I believe it is not unthinkable to imagine that the two interpretations of Galileo's principle of relativity (or some version of them) could have been formulated even in the seventeenth century.

It does seem *logical* to interpret Galileo's principle of relativity to mean that we cannot discover absolute uniform motion because it does not exist. This interpretation is more logical since if uniform motion in space does not exist it becomes immediately clear why it cannot be detected. But if there is no absolute uniform motion, it follows that there is no uniform motion at all because absolute uniform motion is motion in space and if absolute uniform motion does not exist it means that uniform motion in space does not exist either. In such a case the only motion possible would be accelerated motion since such a motion can be detected due to the existence of inertial forces. (We all know this from our everyday experience; the question of the reality of inertial forces will be discussed specifically in Chap. 10.) In terms of being detectable, accelerated motion is absolute, which seems to imply that it is motion with respect to an absolute space. Such a conclusion appears to support the traditional research team since, if absolute accelerated motion exists, absolute uniform motion should also exist – we all see that *both* accelerated and uniformly moving bodies move in space. We shall return to the question of whether absolute accelerated motion requires an absolute space in Chap. 4, but let us now concentrate on the problem with absolute uniform motion which the radical research team is facing. The conclusion that uniform motion in space does not exist is an apparent paradox because hardly anyone would deny that objects move uniformly *in* space.[5]

At this moment the traditional research team may triumph. Such an apparently obvious contradiction with what we observe undoubtedly appears to mean that the second interpretation of Galileo's principle of relativity is wrong. The radical research team, however, may fight back by pointing out a basic paradox which arises from the first interpretation: why, for instance, we cannot discover the motion of a ship *with respect to space* – we all see that the ship moves *in space*. The first interpretation of Galileo's relativity principle – that absolute

[5] One may argue that, since any motion of an object, according to our everyday experience, is always determined with respect to other physical objects (because space is not tangible), what matters is motion with respect to other objects (recall the relationist position), not motion with respect to space. Such a claim, however, is an operationalist one, since it is based on what we can *measure*. We can measure the motion of an object with respect to other objects, but not with respect to space, despite the fact that the object is clearly moving in space. Operationalism may work fine when a science is *applied*, but it is of little or no help in cases when fundamental questions are asked. The question both research teams are asking is precisely *why* we cannot measure the uniform motion of an object in space.

uniform motion exists but cannot be detected – does lead to the following paradoxical situation: we can say that an object moves *in space* (since we cannot deny the obvious), but it would be incorrect to say that the object moves *with respect to space*[6] (since 'with respect to space' implies that the object's motion can be measured as in the case of an object moving with respect to another object). This situation does appear to be paradoxical because our everyday experience tells us that 'motion *in* something' is equivalent to 'motion *with respect to* something'. That is why, no matter whether space is considered to be an entity or a collection of relations, it follows that motion *in* it is motion *with respect to* it. As we will see shortly, this is one of the many instances when our everyday experience has provided us with reliable knowledge.

As each team faces paradoxes that appear insurmountable, we can imagine that they will be hard at work. The best strategy that the radical research team can employ is not to give up when such obvious paradoxes are reached. They can even develop a special approach toward such paradoxes – if something appears obviously wrong it may be assumed that it is too obvious to be wrong. They will soon realize that when direct contradictions with our everyday experience and knowledge are discussed, a necessary condition for successful analysis is to try to turn off their common sense and define everything in terms of logical statements. After they have formulated all logical possibilities, they can turn their common sense on again. Such a strategy, which requires a lot of dedication to achieve, may help them realize that a single implicit assumption – that there is just *one* space – causes not only the paradox they face – that there is no uniform motion at all if there is no absolute uniform motion – but also the paradox reached by the traditional team – that it is correct for an object to move *in* space but not *with respect to* space.

If there are indeed more spaces, it becomes immediately clear that uniform motion can still exist but it is not an absolute uniform motion, which is by definition uniform motion with respect to the absolute space, meaning with respect to the *single* space. An object can be at rest in one space but uniformly moving at different speeds in other spaces. The paradox that the traditional research team faced is also

[6] A similar situation exists even today in the framework of the standard presentations of special relativity – we can say that an object moves in space (again because we cannot deny the obvious), but it is not correct to say that the object moves with respect to space since it would mean that the object is in absolute motion.

immediately resolved – when an object moves uniformly in space it moves in *any* space (except the one in which it is at rest), but we cannot say that the object is moving uniformly with respect to space since we must specify with respect to *which* space the object is moving. So, by realizing the implicit assumption that there exists just one space, which caused both paradoxes, the radical research team did a favour to their colleagues from the traditional team.

The resolution of the two paradoxes is spectacular and it does demonstrate to both teams the power of the method of exploring the internal logic of fundamental ideas. However, to have a logical resolution of a paradox is one thing, but to be able to translate that resolution in terms of our everyday experience is quite another. It is easy to say that there are more spaces, but we all perceive the physical objects as existing in a single three-dimensional space. That is why it is really difficult, to say the least, to consider seriously the existence of more three-dimensional spaces.

The radical research team has already gained sufficient experience in dealing with such apparently obvious contradictions with what we perceive. So, let us see what we really perceive. We see objects at different distances from us and believe that all of them exist in the *same* three-dimensional space. However, in 1676 Römer measured that light was propagating at a *finite* speed, which demonstrated that we see only *past* images of objects, since it takes some time for the light from those objects to reach our eyes. As a three-dimensional space is defined as all space points existing *simultaneously* at a given moment of time, it follows that objects which we see at different distances from us have, in fact, existed in different three-dimensional spaces belonging to different moments of time (Fig. 3.1).

Therefore what we 'perceive' through the objects we see is *not* a space at all – the *same* three-dimensional 'space' (the inclined lines in Fig. 3.2) we 'perceive' is in fact a set of fragments from *different* three-dimensional spaces corresponding to different moments of time. As Fig. 3.2 clearly shows, our belief that we perceive objects occupying the *same* three-dimensional space is wrong.

But the different three-dimensional spaces corresponding to different moments of time are not the three-dimensional spaces that led to the resolution of the two paradoxes, since that resolution required that at any moment there should exist more than one space. At this stage of the analysis it cannot be immediately concluded how the assumption of different spaces should be understood. There exist two logical possibilities:

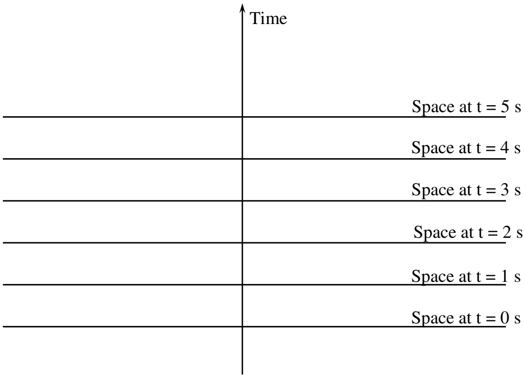


Fig. 3.1. Three-dimensional spaces corresponding to different moments of time. Only one spatial dimension is shown in the figure

- the different spaces are *non-coinciding*, which implies that these spaces are three-dimensional cross-sections of a space of at least four dimensions (as different lines, which are one-dimensional spaces, are one-dimensional cross-sections of at least a two-dimensional space),
- the different spaces coincide (in the sense that there is no need for a space of extra dimensions) but are in translational motion.

What is important, however, is that we do not 'perceive' the *same* three-dimensional space. And this is a crucial achievement since it can provide us with a hint on how different three-dimensional spaces may exist. If the 'space' we 'perceive' through the objects we see consists

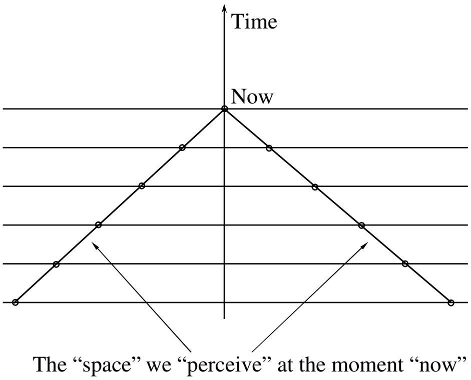


Fig. 3.2. The 'space' we 'perceive' at the moment 'now' consists of fragments (the *small circles*) from different three-dimensional spaces corresponding to different moments of time

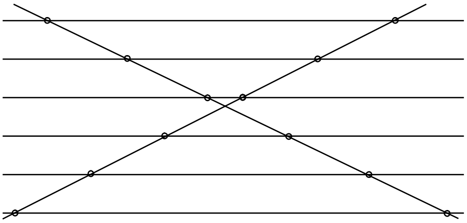


Fig. 3.3. A logical possibility based on the fact that what we 'perceive' at the moment 'now' is not the *same* three-dimensional space – two three-dimensional spaces consisting of fragments from different three-dimensional spaces which correspond to different moments of time

of fragments from three-dimensional spaces corresponding to different moments of time, then it is a logical possibility that different three-dimensional spaces can be constructed in the same way – as consisting of fragments from three-dimensional spaces belonging to different moments of time (the inclined lines in Fig. 3.3).

A possible objection to this logical possibility is that the horizontal lines in Fig. 3.2 do not represent *different* three-dimensional spaces; it is simply the *same* three-dimensional space at different moments of time. The radical research team is unlikely to be impressed by such an objection since it is dictated by our common sense, which is not a reliable authority in the eyes of the scientists from that team; as they advance in their analysis of the second interpretation of Galileo's principle of relativity, their reliance on our common sense gradually diminishes. And indeed the radical team of researchers can advance an immediate argument: if the analysis of Galileo's relativity principle requires that there exist more three-dimensional spaces, then it is hard to accept arguments presupposing the existence of a *single* three-dimensional space. The radical researchers will not accept the argument that there exists a single three-dimensional space which is the same at the different moments of time since it is the very question they have been trying to answer. If at the different moments of time there exists the same three-dimensional space, then the assumption of many spaces may be interpreted to mean that at every moment there exist many *coinciding* three-dimensional spaces in translational motion, which could have been mistakenly regarded as the same three-dimensional space because, due to their complete overlapping, there is no need for an extra-dimensional space. However, the other interpretation of the many-spaces assumption – that those spaces do not coincide – implies that there are different three-dimensional spaces belonging to the different moments of time as seen in Fig. 3.3. As at this moment the members of the radical research team do not know which interpretation of the many spaces assumption is consistent with the non-existence of absolute uniform motion, they obviously cannot accept the common view that the same three-dimensional space belongs to different moments of time.

There is one more consequence of the realization that the implicit assumption of just one three-dimensional space may be responsible for the difficulty in understanding the physical meaning of Galileo's principle of relativity. The fact that space at a given moment of time is defined as all points that correspond to that moment means that the three-dimensional space consists of all points that exist *simultaneously* at this moment. If different three-dimensional spaces are interpreted to mean non-coinciding spaces, then it follows that more three-dimensional spaces imply different sets of simultaneously existing points. Therefore, if space is not absolute, in this interpretation of the many-spaces assumption, simultaneity is not absolute either. In the other interpretation of that assumption (regarding the different spaces as coinciding), simultaneity remains absolute since the different sets of points, constituting different spaces, correspond to the same moment of time. However, as we will see later, different spaces in relative motion do imply different sets of simultaneous events, and this means that simultaneity is not absolute.

Our analysis has demonstrated that there would be no absolute uniform motion if there were more three-dimensional spaces. However, that conclusion tells us nothing about how many three-dimensional spaces there are, what determines their existence and number, and whether the assumption of many spaces implies an extra-dimensional space.

In order to gain some additional insight into these questions, let us join the traditional research team whose scientists are trying to answer the question as to why we cannot detect absolute uniform motion. As traditional and more conservative thinkers, these scientists are not impressed by the resolution of their paradox with the help of the radical idea of many spaces. What is even worse is that the assumption of three-dimensional spaces in relative motion puts an end to the idea of absolute uniform motion (as motion in *the* space) and therefore threatens their basic belief in the existence of absolute uniform motion. That is why the members of the traditional research team have been thinking of using different kinds of experiments to try to detect the absolute uniform motion of objects. They have essentially been considering a

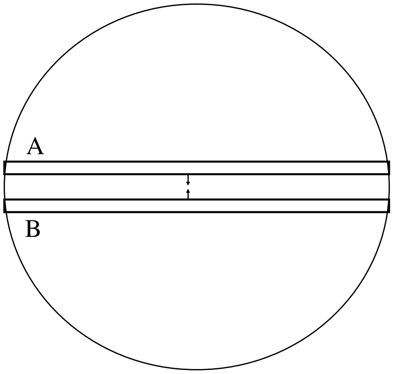


Fig. 3.4. Thought experiment with two spacecraft A and B at rest with respect to each other. A light sphere, emitted from the middle points of A and B (where the heads of the *small arrows* meet), reach the end points of the spacecraft simultaneously

reformulation of the first interpretation of Galileo's principle of relativity – from "absolute uniform motion exists but cannot be discovered" to "absolute uniform motion exists but cannot be detected with the type of experiments which Galileo discussed and performed".

Since Galileo studied only mechanical experiments, the traditional research team decides to see whether experiments involving light can help them discover the absolute motion of the Earth. Such a thought (Gedanken) experiment was not unthinkable in the last quarter of the 17th century, since Römer's experiment in 1676 had already determined that the speed of light was finite. In fact, Galileo himself suspected that light propagated at a finite speed and even contemplated an experiment to measure the light speed by using lanterns on distant hills. Scientists in the 17th century could have used ordinary ships in their thought experiment, but let us consider two spacecraft A and B[7] (Fig. 3.4).

Initially the spacecraft are at rest with respect to each other. The observers in A and B perform the following experiment. An electric spark between the middle points of A and B (the two small arrows in Fig. 3.4 representing two pieces of wire) gives rise to a spherical light wave which simultaneously reaches the end points of the two spacecraft (the circle in Fig. 3.4). Here it is clear that the traditional scientists

[7] While in relative motion the two spacecraft constitute two inertial reference frames; any body moving by inertia (with uniform velocity) can be regarded as an inertial reference frame.

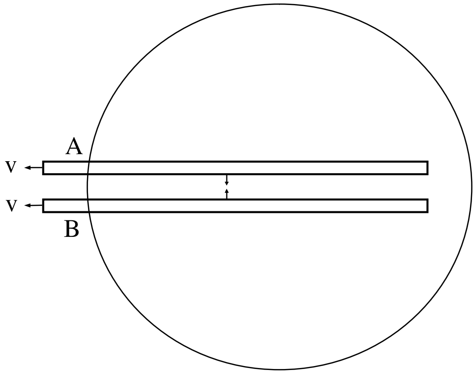


Fig. 3.5. Thought experiment with two spacecraft A and B at rest with respect to each other, but moving in the absolute space. A light sphere emitted from the middle points of A and B does not reach their end points simultaneously. The light sphere has already reached the rear end points of A and B and is now chasing their front end points

explicitly assumed that A and B were in a state of absolute rest (or at rest in the absolute space). Only in this case would the propagating light sphere reach the end points of A and B simultaneously; otherwise, if the spacecraft were moving together in space, the light sphere would reach their rear end points first as shown in Fig. 3.5.

Assume now that B goes away, turns back and starts to move toward A at a constant velocity. At the moment the two spacecraft momentarily coincide, an electric spark again produces an expanding spherical light wave. The observers in A will determine that, as the light wave propagates toward the rear end point of B, B moves in the opposite direction and the light wave will reach the rear end point of B before reaching its front end point, since according to the A-observers light will travel less distance to the rear point of B (Fig. 3.6).

Now the crucial question is: What will the observers in B determine? If they observed what the people in A see then this would be the end of Galileo's principle of relativity since the observers in B would find a discrepancy between the experiment they performed in Fig. 3.4 and the same experiment when they were moving with respect to A as shown in Fig. 3.6. In the first case the people in B observed that light simultaneously reached the end points of B, whereas in the second experiment they would see that light reached the rear end point of B first. Therefore the B-observers would be able to say that they detected their absolute uniform motion by performing a non-mechanical experiment involving light.

The reader can again join the triumphant traditional research team and try to figure out with them why mechanical experiments have failed to detect absolute uniform motion. But before wishing you luck in this apparently easy task I would like to invite you to continue to follow the analysis of the radical research team, as an exercise in creative and analytical thinking. And bear in mind what the radical researchers reminded their traditional colleagues – that the proponents of the Ptolemaic system similarly regarded the tower experiment as an irrefutable argument against the heliocentric model of the solar system.

No matter how eccentric the scientists of the radical research team might look, they just do not believe that the observers in B will determine what the people of A find. The radical team cannot consider such an option since it would mean that the B-observers would discover their absolute uniform motion – something that does not exist according to this team. Therefore, the people in B should observe exactly the same thing they observed when they were at rest with respect to A: that the light sphere reached the end points of B simultaneously. But there is a surprising price for this requirement – the light sphere can reach the end points of B simultaneously only if the speed of light is *constant* and does not depend on the state of motion of the source or the observer. To have a clearer view of this conclusion, assume that B does see what the people in A see. This would mean that for B light travels faster toward the rear point of B than toward the front point of B (since light would travel the distance from the middle point of the B spacecraft to the rear end point for *less* time). Such an observa-

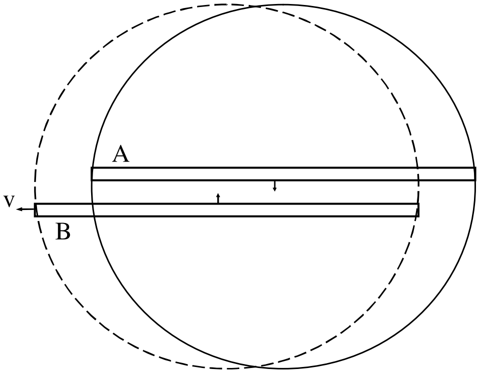


Fig. 3.6. Thought experiment with two spacecraft in relative motion

tion would be consistent with the concept of absolute uniform motion: spacecraft B moves in the absolute space where light propagates at speed c and since B moves to the left at speed v, it will appear to the observers in B that it is the space that moves to the right at speed v and carries the propagating light; this means that for B the speed of light will be $c + v$ since it moves to the right in space and space itself appears to move to the right relative to B. By the same argument the speed of light travelling toward the front end point of B with respect to B will be $c - v$. With this velocity, light will arrive at the rear end point of B first, as determined by the observers in B as well, which would explain why both A and B would observe the same thing.

It should be stressed that the constancy of the speed of light follows from the second interpretation of Galileo's relativity principle. If the speed of light were not constant as we saw in the preceding paragraph, it would mean that absolute uniform motion could be discovered.

The constancy of the speed of light is indeed quite counter-intuitive, but it follows directly from the second interpretation of Galileo's principle of relativity and should be analyzed to understand its physical meaning. This can be done by recalling the fact that the radical research team has already arrived at the conclusion that there are more three-dimensional spaces, not just one. If the observers in A and B have two different three-dimensional spaces, then the constancy of the speed of light is not a puzzle any more – every observer measures the speed of light in his three-dimensional space and for this reason always finds that light propagates at the same speed c there.

Another problem that arises from the thought experiment in Fig. 3.6 is the disagreement of the A- and B-observers over whether or not light arrives *simultaneously* at the end points of B. Our common sense tells us that such a disagreement is an obvious paradox. However, when we again examine our initial assumptions we will discover that we implicitly assumed that simultaneity is absolute. How do we know that? Do we have any arguments to support such an assumption? Not only will the researchers of the radical research team not be surprised at the disagreement between A and B over what is simultaneous but they will be quite excited since the thought experiment offered by the traditional team can help them eliminate one of the two possible interpretations of the many-spaces assumption. As we have seen, that assumption already implies that simultaneity will not be absolute if the different spaces form an angle with one another which requires at least four-dimensional space; simultaneity would be absolute only if there existed a single three-dimensional space or there were different

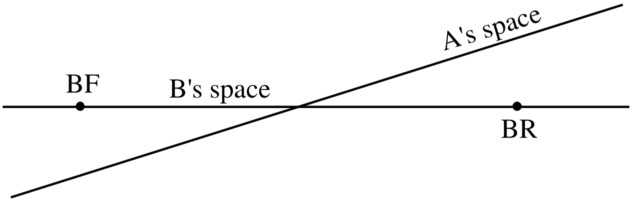


Fig. 3.7. Simultaneity is not absolute for the observers in two spacecraft A and B in relative motion

three-dimensional spaces in relative motion, but they were overlapping and therefore there was no need for an extra-dimensional space to accommodate many spaces. Therefore the thought experiment depicted in Fig. 3.6 does eliminate the interpretation that many spaces can move relative to one another translationally in such a way that the three-dimensionality of space is preserved.

Let us now see what additional information the radical research team can obtain from an analysis of Fig. 3.6. The fact that for the observers in spacecraft A in Fig. 3.6 the propagating light sphere reaches the rear end point of spacecraft B (event BR) first and then its front end point (event BF), whereas the observers in B claim that the two events should occur simultaneously, can be depicted as in Fig. 3.7.

The radical research team can easily realize that, having different spaces, the A- and B-observers must have different times as well. What helps them reach this conclusion is the discussion that different spaces correspond to every moment of time as depicted in Fig. 3.1. Applied to the A- and B-observers this means that the two different sets of spaces corresponding to the different moments of every observer define *different times* for both observers, as shown in Fig. 3.8.

Now the radical team will be able to realize fully the profound consequences of Galileo's principle of relativity. Galileo's observations and arguments that motion cannot be detected by performing mechanical experiments lead to the conclusion that there is no absolute uniform motion, and this in turn is only possible if there are more three-dimensional spaces. The analysis of the thought experiment proposed by the traditional research team to detect absolute uniform motion with the help of light signals showed that absolute uniform motion can be detected only if the speed of light is not constant and simultaneity is absolute. In other words, if absolute uniform motion does not exist (and therefore cannot be detected), there are two immediate consequences:

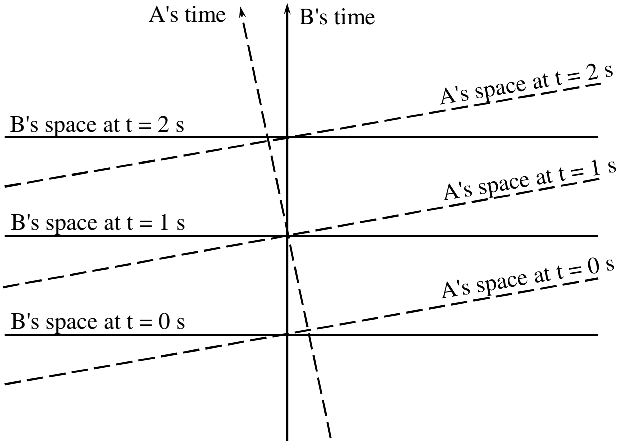


Fig. 3.8. Relativity of simultaneity implies different spaces and different times

- the speed of light is constant,
- simultaneity is not absolute.

This result helped the radical research team to conclude that the assumption of the existence of different three-dimensional spaces requires a four-dimensional space with time as the fourth dimension, because non-coinciding three-dimensional spaces can exist only as three-dimensional cross-sections of at least four-dimensional space. However, such a conclusion appears to imply that the world itself is four-dimensional. This becomes evident when the question of the dimensionality of the world is explicitly asked. The members of the radical research team recall that, since the time of Aristotle, the world has been regarded as three-dimensional (see [4, pp. 9–10]).

But what is the world? What our senses are telling us is that everything that we see at the present moment is what exists. However, after Römer had determined that the speed of light was finite, it became clear that what we see at the moment 'now' existed some time ago. Since then the common-sense view (called *presentism*) has been that the world is *the present* defined as everything that exists *simultaneously* at the present moment. What is essential for the radical research team is that the present – the three-dimensional world at the moment 'now' – is defined in terms of *simultaneity*. This means that relativity of simultaneity immediately affects the presentist view – having different sets of simultaneous events, the observers in spacecraft A and B in Fig. 3.6 have different presents and therefore different three-dimensional worlds. However, this is only possible if the world is four-dimensional with time as the fourth dimension. Otherwise, if the world were three-dimensional, the observers in A and B would have a common three-dimensional space and therefore a common set of simultaneous events, which means that simultaneity would be absolute.

A four-dimensional world, however, is completely counter-intuitive. As the time dimension in a four-dimensional world is *entirely* given, like the three spatial dimensions, nothing happens in such a world – it is a frozen, timelessly existing world. At this point the radical team faces its greatest challenge – how can one seriously claim that the external world is so dramatically different from what our senses tell us about it? Some members of this team may even find that they have had quite enough of pursuing only the radical options in the analysis of Galileo's principle of relativity and may feel that it is more rational to join the traditional research team. So let us see what problems they will have to deal with when they start their research with their traditional colleagues.

This team postulated that absolute uniform motion did exist but for some reason we could not detect it. In order to see whether light can be used to detect absolute motion, they proposed the thought experiment represented in Figs. 3.4 and 3.6. Unlike the radical research team, the members of the traditional team were convinced of the obvious – that the observers in both spacecraft should observe that the propagating light sphere reaches the rear end point of spacecraft B first and after that B's front end point. Therefore they naturally arrive at the conclusion that absolute uniform motion does exist and can be detected by experiments involving light. Everything looks so simple and self-evident. There is no need to invoke such extreme and exotic hypotheses about many three-dimensional spaces and a four-dimensional frozen space, or to deny the absoluteness of simultaneity. Moreover, the researchers from this team see an insurmountable contradiction in the radical team's interpretation of the thought experiment shown in Fig. 3.6: if simultaneity is relative then, as seen in the figure, there will be *two* propagating light spheres each centered at the middle of each spacecraft; but, by definition, there was *one* light signal and therefore there should be one light sphere. The radical team's explanation that each of the spacecraft has its own three-dimensional space and for this reason there are two spaces in which two light spheres propagate is dismissed by the traditional team as an example that one can explain everything with the help of ad hoc hypotheses. The traditional scientists find the explanation that a *single* light signal can split into *two* propagating light spheres a complete fantasy.

As the members of the traditional research team are also good researchers, they decide, while awaiting their colleagues from the experimental physics group to perform the experiment depicted in Fig. 3.6 and to confirm experimentally the existence of absolute uniform motion, to analyze the only open question – why Galileo had failed to detect absolute motion with mechanical experiments.

The answer seems obvious – as we have seen in Chap. 2 it is the inertia of the objects involved in all mechanical experiments designed to test absolute uniform motion that is responsible for the null results. This may be a perfect explanation for the members of the traditional research team, but the newcomers from the radical team, who are accustomed to asking more 'why' questions, may remind their new colleagues that the very existence of inertia demonstrates that *the* space offers no resistance to the uniform motion of bodies. Then a question that follows naturally is: Why do bodies encounter no opposition from space (regarded as an entity) during their motion in it?[8]

Now the researchers from the traditional team will find themselves in a non-traditional role. To answer this question they have to choose between two extreme assumptions:

- space is indeed an entity (some kind of medium – we all see that there is 'something' between the stars in the night sky) which, however, unlike any other entity, offers no resistance to uniform motion (but miraculously resists accelerated motion),
- space is not an entity and for this reason offers no opposition to motion.

It would have been quite interesting to listen to the arguments of the traditional team designed to persuade the other traditional scientists, not involved in the research on the physical meaning of Galileo's principle of relativity, that space is a non-entity. But this period would not last for too long. Experiment – the ultimate judge – would rule against the traditional team. At the turn of the 20th century the Michelson–Morley experiment would show that experiments involving light cannot detect absolute uniform motion either.

However, let us not wait for the Michelson–Morley experiment, but return instead to the remaining members of the radical research team

[8] The former members of the radical research team may even start to analyze the possibility that the existence of inertia may mean that there is no absolute uniform motion.

whose intellectual curiosity helped them overcome the initial shock from the realization of what a four-dimensional world is. They are also about to receive an unexpected present from their former colleagues who joined the traditional research team. The former radical researchers visited their colleagues of the radical research team and told them about their suspicion that the existence of inertia may be a manifestation of the non-existence of absolute uniform motion. They informed the radical researchers of the traditional team's belief that it is inertia that prevents us from detecting the uniform motion of objects in space. It did not take much time for the present and former radical researchers to rule out such an explanation for at least two reasons. First, they did not believe that Nature conspires against us by inventing inertia to prevent us from detecting uniform motion in space. This team believed that if something exists it should be detectable in principle. So, if absolute uniform motion existed we should be able to discover it. That is why for them it is unlikely that inertia conceals absolute uniform motion; they rather believe that inertia results from the fact that there is no such motion.[9] Second, the existence of inertia cannot resolve the paradox the traditional research team faced – why is it impossible to say that an object, which moves uniformly *in* space, moves *with respect to* space as well?

As inertia does not explain the unsuccessful attempts to detect absolute uniform motion with mechanical experiments, it follows that if such a motion existed it should be detected even with the type of experiments Galileo performed. This conclusion provides strong and independent support for the second interpretation of the physical meaning of Galileo's principle of relativity – that there is no absolute uniform motion – and makes the members of the radical research team more certain than ever that the direction they have been pursuing will lead them to the truth. Now they can return to the worrying questions posed by the conclusion they reached that the world is four-dimensional. Naturally, they take some time for reflection on what reality is and realize that the disturbing world view, following from the second interpretation of Galileo's principle of relativity which denies the existence of absolute uniform motion, is reminiscent of the Eleatic view, according to which the true reality is an eternal existence. The next step of the radical team is perfectly traditional. They would like

[9] As we will see in Chaps. 4 and 10 this really appears to be the case – the very existence of the inertia of an object (moving with uniform velocity on its own and resisting its acceleration) is most likely a manifestation of the four-dimensionality of the world.

to see whether the four-dimensionalist view contradicts our experience based on the fact that what we perceive are three-dimensional images[10] of the external world. Their analysis clearly shows that our perception has two possible explanations: the three-dimensional images we realize are either caused by a three-dimensional external world or these are three-dimensional images of a multi-dimensional world.

Once the radical research team has realized the two possible explanations of our perceptual experience, it becomes evident that the four-dimensionalist view will contradict our experience only if we can provide a proof that the second explanation of what we perceive is wrong. After an intense discussion, the researchers of the team not only fail to find such a proof, but come to the conclusion that what they have been doing from the beginning of their analysis of Galileo's principle of relativity may provide a proof of that explanation. Now nothing can stop them from realizing the most important ideas of the theory of relativity. Once they have discovered that simultaneity is not absolute, they find out what its physical meaning is and conclude that, having different sets of simultaneous events, different observers in relative motion will have different three-dimensional spaces, which is only possible if reality is a four-dimensional world. (If reality were a three-dimensional world, there would exist just one three-dimensional space.) Then all consequences of the theory of relativity can be easily obtained as we will see in the next chapter.

3.2 On the Two Postulates of Special Relativity

If we look at the path of the radical team to the four-dimensional formulation of relativity, we will see that it is more powerful, consistent, and convincing than that presented in the standard formulation of special relativity. Let us start with the original two postulates given by Einstein in his 1905 paper [11]:

> Examples of this sort, together with the unsuccessful attempts to discover any motion of the earth relatively to the 'light medium', suggest that the phenomena of electrodynamics as well as of mechanics possess no properties corresponding to the idea of absolute rest. They suggest rather that, as has already been shown to the first order of small quantities, the same laws of electrodynamics and optics will be valid for all frames of

[10] In fact we perceive two-dimensional images which are processed by the brain in order to become three-dimensional.

> reference for which the equations of mechanics hold good. We will raise this conjecture (the purport of which will hereafter be called the 'Principle of Relativity') to the status of a postulate, and also introduce another postulate, which is only apparently irreconcilable with the former, namely, that light is always propagated in empty space with a definite velocity c which is independent of the state of motion of the emitting body. These two postulates suffice for the attainment of a simple and consistent theory of the electrodynamics of moving bodies based on Maxwell's theory for stationary bodies. The introduction of a 'luminiferous ether' will prove to be superfluous inasmuch as the view here to be developed will not require an 'absolutely stationary space' provided with special properties, nor assign a velocity-vector to a point of the empty space in which electromagnetic processes take place.

Einstein merely postulated that the idea of absolute space was not necessary for the new theory. As a result, any questions about absolute uniform motion and absolute rest were regarded as meaningless. Einstein did not answer the fundamental question of *why* there is no absolute space and therefore no absolute uniform motion and absolute rest, but at least he explicitly stated that special relativity did not need absolute space. In this respect the situation has not changed in the presentations of special relativity that came after Einstein. Take for example a typical formulation of the two postulates of special relativity (see for instance [13–17]):

- The laws of physics are the same in all inertial reference frames.
- The velocity of light in vacuo (c) is the same in all inertial frames.

There is not a single word about the impossibility of detecting the uniform motion of an object *in* space. I believe this question should be *explicitly* addressed in the postulates of special relativity since we all see that objects are moving *in* space. Then it is a valid question to ask why we can say that an object moves *in* space, not *with respect to* space. In fact, the first postulate, which is sometimes called the generalized principle of relativity, is essentially saying that absolute uniform motion cannot be discovered by *any* experiments. That is why the form of the laws of physics in all inertial reference frames is the same; otherwise, if an observer detected that the form of a law was different in his inertial reference frame he would be able to claim that he had discovered his absolute uniform motion.

Note that the first postulate implicitly states that absolute uniform motion cannot be discovered, no matter what kinds of experiments are carried out. And this is done on the basis of only two types of experiment – mechanical and electromagnetic. But what about other types of experiment involving, say, weak and strong interactions? It is true that we have good reasons to believe that any experiments will give null results. But what is important is that we do not know *why* absolute uniform motion cannot be detected and for this reason cannot be certain that such a detection is *in principle* impossible.

By contrast, the members of the radical research team have directly addressed the questions of absolute space and absolute uniform motion. They interpreted Galileo's principle of relativity by explicitly assuming that absolute uniform motion does *not* exist. The members of this team tried to understand the *experimental fact* that the obvious uniform motion of a ship in the water, in the air, but *in* space as well could not be detected by mechanical experiments. They believed that there should be an explanation for this fact. And their analysis did lead them to the conclusion that uniform motion in space could not be detected because there are more spaces, which means that absolute uniform motion does not exist.

Once it has been concluded that absolute uniform motion does not exist, it immediately follows that no matter what kind of experiments one carries out, they will all give null results – one cannot detect what does not exist. That is why a postulate explicitly stating that there is no absolute uniform motion (based on the experimental evidence that such a motion cannot be discovered) is more powerful than the so called generalized principle of relativity. A postulate which is even more powerful is the one given in Minkowski's formulation of special relativity – that the world is four-dimensional. And indeed, as we have seen, the profound physical meaning of the non-existence of absolute uniform motion is that the world is four-dimensional, which in terms of our ordinary three-dimensional language means that observers in relative motion have different three-dimensional worlds which are three-dimensional cross-sections of this four-dimensional world.

The other immediate consequence of the interpretation of Galileo's principle of relativity that there is no absolute uniform motion is the constancy of the speed of light – the speed of light must be the same in all inertial reference frames because if this were not so, absolute uniform motion would be discovered as we have seen in the discussion of the thought experiment shown in Fig. 3.6. Therefore the method of exploring the internal logic of fundamental ideas followed by the radical research team revealed that the constancy of the light speed is a consequence of the relativity principle and not an *independent* postulate. Another obvious advantage of this method is its potential for obtaining ground-breaking results such as the surprising conclusion that the non-existence of absolute uniform motion requires the existence of many three-dimensional spaces and therefore the existence of a four-dimensional world.

The fact that observers in relative motion have different three-dimensional spaces is not clearly addressed in the standard presentations of special relativity. In 1908 Minkowski noticed that [9, p. 83]: "Neither Einstein nor Lorentz made any attack on the concept of space." He was the first to point out that the idea of many spaces is inevitable for a true understanding of special relativity [9, pp. 79–80]:

> We should then have in the world no longer *space*, but an infinite number of spaces, analogously as there are in three-dimensional space an infinite number of planes. Three-dimensional geometry becomes a chapter in four-dimensional physics. Now you know why I said at the outset that space and time are to fade away into shadows, and only a world in itself will subsist.

Unfortunately, Minkowski's analysis does not seem to have been fully appreciated. Such a tendency is evident even in Sommerfeld's notes on Minkowski's paper [12]:

> What will be the epistemological attitude towards Minkowski's conception of the time–space problem is another question, but, as it seems to me, a question which does not essentially touch his physics.

One of the major results of the method employed by the radical research team is the realization that special relativity is *logically* contained in Galileo's principle of relativity. This is so because, as we have seen, this principle logically had only two interpretations and one of them directly leads to the basic ideas of the theory of relativity and to its mathematical formulation, as we will see in the next chapter. The only additional assumption that was used is the experimental fact that the speed of light is finite.

3.3 A Lesson from a Delayed Discovery

The realization that Minkowski's four-dimensional formulation of special relativity is logically contained in Galileo's principle of relativity naturally leads to the question of whether special relativity could have been discovered earlier. I do not intend to speculate on when this could have happened. But I hope this chapter has convinced you that special relativity is indeed a delayed discovery. Then the next unavoidable question is whether future discoveries may also be delayed.

One can arrive at two views on why there have been no recent breakthroughs (of the type of special and general relativity and quantum mechanics) in fundamental science:

- there is not enough experimental evidence to make a breakthrough possible,
- all conditions necessary for a new physical theory, for example, are present, but researchers have been failing to process successfully the existing theoretical and experimental evidence.

The analysis in this chapter shows that the second possibility cannot be excluded. At first sight, this should be no reason for concern. And indeed in many cases delayed discoveries are just that – delayed discoveries and nothing more. What is disturbing, however, is that in some extreme cases delayed breakthroughs in biology or physics, for instance, may turn out to threaten the very existence of the human race.[11]

In order to try to reduce the likelihood of future discoveries being delayed, we have to find a way to make the quest for scientific knowledge more effective by developing research strategies and educating a new generation of creative researchers. A helpful step in this direction might be an international development of a compulsory university course on exploring the internal logic of fundamental ideas as a shorter path to novel ideas and scientific discoveries. Versions of such a course could be developed by scientists working in different fields, who can provide case studies of delayed discoveries in their own fields.

[11] It is not unthinkable to imagine that a wandering planetary system is on a collision course with our solar system. We have to either accept our fate and spiritually prepare for the journey in the non-being or embark on a desperate research project to study the mechanism of inertia and gravitation with the hope that we may learn how to control them. A success might allow us to take the Earth with us and leave our doomed solar system; launching local 'suns' in orbit above the Earth would probably not be as challenging as converting the Earth into a gigantic spacecraft. Obviously, in such a situation a discovery is either made in time or there will be no delayed discovery. Another doomsday scenario involving genetics appears even more probable.

3.4 Summary

In this chapter we have examined the physical meaning of Galileo's principle of relativity and arrived at the conclusion that the null result of Galileo's experiments implies that absolute uniform motion does not exist. Our analysis showed that the non-existence of absolute uniform motion entails the existence of more three-dimensional spaces which in turn is possible only in a four-dimensional world. The results of this analysis demonstrate that the four-dimensional formulation of special relativity is logically contained in Galileo's principle of relativity.

4 Relativity in Euclidean Space and in Spacetime

> The whole universe is seen to resolve itself into ... world-lines, and I would fain anticipate myself by saying that in my opinion physical laws might find their most perfect expression as reciprocal relations between these world-lines.
>
> H. Minkowski [9, p. 76]

As we have seen in Chap. 3 the most disturbing consequence of the analysis of Galileo's principle of relativity carried out by the radical research team was that the world is four-dimensional. But so far the results from their analysis have not encountered any immediately obvious contradiction with the existing experimental evidence. It is natural for the researchers of this team to want at this stage to inform the scientific community of their ground-breaking results. Although they are quite aware of what a difficult task this would be, they know that the best way to convince the skeptical and conservative scientific community is to let the ultimate judge – the experimental evidence – do the job for them. They therefore start working on a mathematical formalism describing the four-dimensional space with time as the fourth dimension.

Again, the radical research team follows the best strategy in times of crisis, when something so obvious – such as absolute motion – leads to contradictions, especially with the experimental evidence: they try to ignore the protests from their common sense and just follow the internal logic of the idea that absolute motion does not exist. They know they might be wrong, but firmly believe that it is worth seeing what predictions that idea will lead to. And as always, it will be experiment that has the last word.

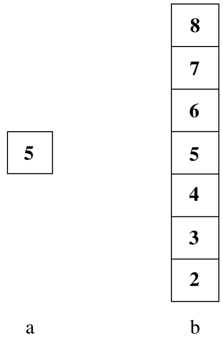


Fig. 4.1. (**a**) A digital clock existing only at the present moment. (**b**) The worldtube of a digital clock in spacetime

4.1 Spacetime

The first step toward mathematical description of the four-dimensional space is to define the correspondence between geometrical and physical objects. As it does not matter what name we will give to the four-dimensional space let us use the current terminology – *spacetime* or *Minkowski spacetime*; Minkowski himself called it the world. We will refer to the spacetime points as *world points* or *events*. As an event is defined by four numbers, three of which give its location in space whilst the fourth tells us at what moment the event occurs, it is important to realize that the concept of event is the basic element of spacetime and that its meaning here differs from its ordinary meaning. The building block of spacetime – the event – is defined as a three-dimensional object, a field point, or a space point at a given moment of time. This definition becomes clear when one takes into account the fact that the fourth dimension of spacetime – the time dimension – is entirely given like the three spatial dimensions. Otherwise, if it were not given at once, there would be no way for spacetime to be four-dimensional. With this in mind it is not so difficult to realize that the whole history of every physical object is entirely given in spacetime as a four-dimensional object. If the physical object is a small point-like particle, that four-dimensional object is just a line[1] in spacetime called a *worldline*. As the worldline of a particle consists of events, *each event is the particle at a given moment of its history in time*.

[1] The realization that a particle is a line in a four-dimensional space does not require any modern knowledge. In 1884 C.H. Hinton explained that ordinary particles will be threads in a four-dimensional space [18, 19].

The radical research team will overcome the temptation to regard the particle's worldline as similar to the particle's trajectory. The worldline cannot be thought of as a line of only one event containing the particle, whereas all other events are empty. If this were the case, the worldline would not be a line in spacetime but would be reduced to just one event – the event that contains the particle. To demonstrate why this is so, consider a digital clock instead of a particle. As it is a spatially extended object, it will be a *worldtube* in spacetime.

Figure 4.1a depicts the ordinary view of a digital clock – the clock exists as a three-dimensional object only at the moment 'now', say at the fifth second. Figure 4.1b shows the digital clock in spacetime – its worldtube consists of the clock at *all moments* of its history.[2] It is evident from Fig. 4.1b that the clock's worldtube cannot be thought of as consisting of one event[3] which contains the three-dimensional clock at its present moment (the fifth second) with all other events being empty. It is important to realize fully that the worldtube of the digital clock is a four-dimensional entity in spacetime. This means that each of the events comprising the clock's worldtube contains a *different* three-dimensional clock at a given moment of its history. If the worldtube contained just one three-dimensional clock – the one showing the present fifth second – the clock's worldtube would not be four-dimensional and therefore there would be no worldtube at all: what would be depicted in Figs. 4.1a and b would be identical three-dimensional digital clocks existing at the fifth second.

Figure 4.1b clearly demonstrates why an event associated with a physical object is the three-dimensional object at a given moment of its history. That is why one should be careful not to confuse the ordinary meaning of the concept 'event' with its meaning as an element of spacetime. For instance, it is a contradiction in terms to think of *different* events happening with the *same* three-dimensional object because *different* events are, as shown in Fig. 4.1b, *different* three-dimensional clocks that exist at the different moments of the clock's history.

Which of the views represented in Figs. 4.1a and b correspond to the real world is a separate question and, at this stage of the analysis of the radical research team, it is an open one. However, we will see in Chap. 5 that none of the kinematic consequences of special relativity

[2] What is depicted in the figure is not very rigorous in one respect – the screens of the clocks are extended in the fourth dimension; three-dimensional clocks should be represented by horizontal lines only.

[3] In the case of spatially extended objects the events associated with these objects are also extended in the spatial dimensions but not in the time dimension.

would be possible if the worldtubes of the physical objects involved in those effects were not real four-dimensional entities.

Another important piece of information that can be derived from Fig. 4.1b is that the length of the clock's worldtube is time. The length between any two events on it is a time period. The worldtube of any object subjected to a *periodic* change resembles a ruler. For this reason the worldtube of a clock can be regarded as a *time ruler* in spacetime.

Having clarified the meaning of the concept of 'event' in spacetime, let us now continue to use worldlines instead of worldtubes. The worldline of any uniformly moving physical object can be used as the time axis of an inertial reference frame associated with that object. Introducing an inertial reference frame in spacetime allows us to talk about space *and* time since we have chosen a time direction along the worldline of an uniformly moving object and can choose the three-dimensional space to be orthogonal to the object's worldline. However, it does not follow from here that spacetime itself is divided into space and time. We are free to choose another time direction and another three-dimensional space orthogonal to that time direction. We are also free to choose the three-dimensional space in such a way that it does not form a right angle with the chosen time direction. The situation in spacetime is similar to drawing x- and y-axes on a plane – we can choose any directions for the two axes. We can also choose a non-orthogonal coordinate system in which the x-axis is not orthogonal to y. The only difference between Euclidean geometry and the geometry of spacetime is that we cannot choose the time axis of an inertial reference frame in any direction in spacetime. As we will see shortly the very fact that the time dimension is different from the three spatial dimensions of spacetime imposes some restrictions.

The radical research team will now be in a position to see how the issue of absolute motion would look in spacetime, or more precisely, whether the idea of spacetime can provide some additional insight into *why* there is no absolute uniform motion. The researchers from this team have already realized the disturbing price of the assumption of non-existence of absolute motion – this assumption leads to the idea of spacetime where there is no motion at all due to all moments of time being given at once as the time dimension of spacetime. And indeed, if we consider a worldline of a particle in spacetime, the worldline is a monolithic four-dimensional entity which exists timelessly in the frozen world of Minkowski spacetime. So, in spacetime, which the radical team believes is the true reality, there is no motion at all.

However, this team knows well that no matter what our theories are, they should be expressed in terms which reflect the way we perceive the world. As we perceive motion of three-dimensional objects in the three-dimensional space, it is clear that, in order to talk about motion, we have to introduce a reference frame in spacetime which defines a time direction and a three-dimensional space. The particle along whose worldline lies the time axis of the reference frame is at rest in its own three-dimensional space. To ask whether the particle is in a state of absolute uniform motion is meaningless since such a question, as we have seen in Chap. 3, implies the existence of only *one* three-dimensional space, whereas we can introduce an infinite number of reference frames in spacetime and therefore can define an infinite number of three-dimensional spaces there. So when we say that the particle moves *in* space, it is not necessary to specify in which three-dimensional space it moves since it moves in all possible spaces, except in its own space. However, in order to say that the particle moves *with respect to* space obviously requires us to point out with respect to which three-dimensional space it moves; in other words, we should point out with respect to what particle the initial particle moves, since a time axis and a three-dimensional space are associated with the worldline of a physical particle. In such a way, the requirement to specify a particle with respect to which another particle moves in fact answers two questions at once:

- why there is no absolute uniform motion,
- why it is meaningful to talk only about relative motion (about motion with respect to a physical object).

Here it should be stressed again that absolute uniform motion will not exist only if the world is four-dimensional, where one can define many three-dimensional spaces.

Figure 4.2 depicts different relations between worldlines and the motions of the corresponding particles. An inertial reference frame is associated with the worldline of particle A – the time axis of A's reference frame is along the worldline of A (which defines a direction in spacetime) and A's three-dimensional space is orthogonal to A's worldline. The worldline of particle B is parallel to A's worldline which means that the distance between the two particles does not change in time and therefore they are at rest with respect to each other. As the worldlines of particles A and B are parallel they define the same time direction and therefore it does not matter with which particle an inertial reference frame is associated – both particles will

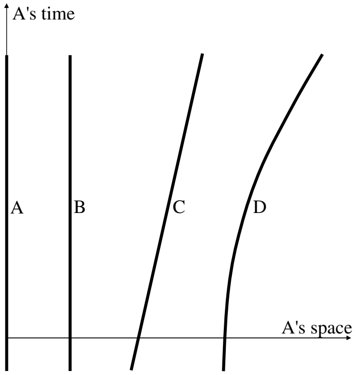


Fig. 4.2. The worldlines of four particles. The worldline of particle A defines the time axis and the three-dimensional space of the inertial reference frame associated with A. Two of the spatial dimensions of A's space are suppressed in the figure

share the same frame and will have common time and common three-dimensional space.

The worldline of particle C is inclined with respect to A's and B's worldlines. This means that C moves relative to A (and B) since the distance between it and A increases with time. Stated another way, C moves uniformly in A's space; but C does not move in its own three-dimensional space which is orthogonal to its worldline (not shown in Fig. 4.2). It is also correct to say that C moves *with respect to* A's space. As shown in Fig. 4.2, in order to talk about motion, there should be two worldlines that form an angle. Therefore, there is a relation between the relative velocity between two particles in space and the angle between their worldlines in spacetime. (The expression for this relation will be obtained in the next section.)

Unlike the worldlines of particles A, B, and C in Fig. 4.2, the worldline of particle D is curved. Looking at the way the distance between A and D changes with time, we easily arrive at the conclusion that D's motion is not uniform; rather, it is accelerated. The comparison of worldlines of particles in spacetime provides an *objective* criterion for the distinction between uniform motion (motion with constant velocity) and accelerated motion: the worldline of a uniformly moving particle is a straight line, whereas the worldline of an accelerating particle is curved.[4] The objective distinction between a straight and a

[4] This criterion holds for the flat Minkowski spacetime. As we will see in Chap. 8, a similar criterion applies in curved spacetime as well.

curved worldline suggests to the radical team that there should be some physics behind this distinction. And indeed they receive another indication that their radical approach may be the right one – uniform motion cannot be detected, whereas accelerated motion can be discovered from within an accelerating reference frame due to the presence of inertial forces there. Also, as we will see in Chap. 8, a non-inertial reference frame, associated with an accelerating particle, can be distinguished from an inertial reference frame,[5] associated with a particle moving with constant velocity (i.e., by inertia), due to the anisotropic propagation of light in the non-inertial frame.

Now the radical research team may claim that they have an explanation for the detectability of accelerated motion – the *deformation* of an accelerating particle's worldtube. However, they still cannot explain the nature of the inertial effects that make the accelerated motion detectable. But they will soon realize that if worldtubes are real four-dimensional objects and since the worldtube of an accelerating particle (which resists its acceleration through an inertial force acting on it) is curved, then it is quite natural to regard the inertial force as originating from a four-dimensional stress in the particle's worldtube which arises when the worldtube is deformed. We will discuss this question in Chap. 10.

The fact that the accelerated motion of a particle can be discovered from within the non-inertial reference frame, in which the particle is at rest, demonstrates that, unlike uniform motion, accelerated motion is absolute. However, it should be pointed out that 'absolute' means detectable, not motion with respect to an absolute space. Often the existence of absolute acceleration is understood to imply the existence of absolute space [20]: "given the definition of absolute acceleration, namely as acceleration relative to absolute space, acceleration is only possible if absolute space exists." But it is clear from the discussion in Chap. 3 that the non-existence of absolute space does not change the fact that accelerated motion is detectable and in this sense absolute. In terms of motion relative to space, acceleration is absolute in the sense that if the worldline of a particle is curved, the particle is accelerating in *all* inertial reference frames, which means with respect to *all* three-dimensional spaces. For comparison, uniform motion is not absolute since a uniformly moving particle is not moving with respect to the three-dimensional spaces of all inertial reference frames, because it is

[5] From now on we will follow the established tradition and will use inertial observers and non-inertial observers as synonyms for inertial frames of reference and non-inertial reference frames, respectively.

at rest in the space of its own inertial reference frame. One may object that an accelerating particle is also at rest in its own space. However, unlike a uniformly moving particle, an accelerating particle does not have its own space. A particle which moves by inertia (with uniform speed) has its own space in the sense that the straight worldline of the particle defines a time direction and its spaces corresponding to different moments of time are conventionally chosen to be perpendicular to the *same* time direction. As shown in Fig. 4.3 the curved worldline of an accelerating particle does not define a given time direction. At any moment of its history the particle has different time directions defined by the tangent lines at the different moments. This means that at any moment a *different* inertial reference frame is associated with the accelerating particle whose time axis is along the time direction at that point. As space is perpendicular to the time direction the particle will have different time directions and different spaces at every moment during the time it accelerates. Physically, this means that at any moment during its acceleration the particle is instantaneously at rest with a different uniformly moving particle and the two particles instantaneously share the same inertial reference frame and therefore the same time and the same space. In Fig. 4.3 the time axes of the inertial reference frames of two observers A and B in relative motion instantaneously coincide with the time directions defined by the tangent lines to the worldline of the accelerating particle at the events M and N. The inertial reference frames of the observers A and B are called instantaneous or comoving inertial reference frames at the events M and N.

The members of the radical research team understand well that if spacetime is real, it is not divided into different three-dimensional spaces and therefore these spaces (and times) are merely *descriptions* of the indivisible four-dimensional spacetime in our three-dimensional language. This is somewhat ironic – one day, the more conservative scientists would regard spacetime as just a description of the real three-dimensional world, but it might turn out that it is the three-dimensional world that is a description of the real spacetime. That is why the radical researchers would prefer to define uniform and accelerated motion of particles, not with respect to space, but in terms of their worldlines which are elements of spacetime and which reflect their experimentally detectable states of motion. If a particle offers no resistance to its motion, it is moving uniformly on its own (i.e., by inertia) and its worldline is a *straight* line in Minkowski spacetime. If a particle resists its motion, it is accelerating and its worldline is

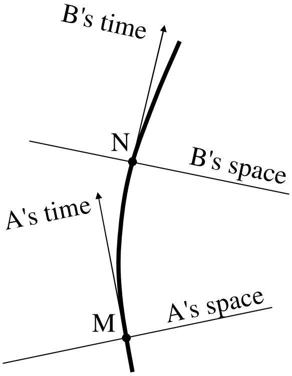


Fig. 4.3. As the worldline of an accelerating particle is curved, at any moment of the particle's time, the tangent to its worldline defines a time direction. This means that the accelerating particle does not have its own time direction and space and at every moment of its time shares the time directions and spaces of different uniformly moving particles with respect to which it is instantaneously at rest. The inertial reference frames of two observers A and B are such instantaneous or comoving inertial frames

curved. The idea of defining uniform and accelerated motion of particles in terms of the shape of their worldlines is also dictated by the following reason. If the radical research team is correct and the concept of spacetime adequately represents the external world, then the *motion* of three-dimensional particles in space does not have an objective analog and is also a *description* of the way we perceive the particles' worldlines. Since there is no motion in space, if spacetime is real, what we perceive as uniform and accelerated motion of particles should be represented by straight and curved worldlines, respectively.

In Chap. 3, we saw that the traditional and radical research teams had completely different views on inertia. The traditional scientists believed that absolute uniform motion exists but cannot be discovered with mechanical experiments due to the inertia of the physical bodies involved in such experiments. For the members of the radical research team, inertia is a *consequence* of the non-existence of absolute uniform motion. And since the non-existence of absolute uniform motion implies a four-dimensional world, it follows that inertia is a manifestation of the four-dimensionality of the world. Such a conclusion is consistent with the link between the shape of the worldtube of a particle and its state of motion. A uniformly moving particle moves on its own and does not resist its motion. However, it is not clear why there is no resistance to the particle's motion and what makes it move on its own. The spacetime picture offers an unexpectedly simple explanation – the particle's worldtube is straight until another worldtube curves it, which means that the particle moves uniformly on its own until another particle prevents it from doing so. As the members of the radical research team regard the particle's worldtube as a real four-dimensional object, they will certainly arrive at the unavoidable conclusion that, like an ordinary three-dimensional rod, a straight worldtube is not deformed and does not resist any deformation; therefore, in three-dimensional language, the particle offers no resistance to its uniform motion. The worldtube of an accelerating particle, however, is curved and it should resist its deformation which means that the particle should resist its accelerated motion. The radical research team will be delighted – as they expected, inertia does appear to be a manifestation of the four-dimensionality of the world:

- a straight worldtube in Minkowski spacetime represents a particle which moves uniformly on its own and does not resist its uniform motion,
- a curved worldtube resists its deformation and represents a particle which accelerates and opposes its acceleration.

The next step in understanding the relations between our perceptual experience and the timelessly existing four-dimensional objects of spacetime the radical research team is likely to take is to examine what the perceived propagation of light in our three-dimensional space looks like in spacetime. Before doing that, however, they need to pay closer attention to the time axis that represents the time dimension. The reason is that they should determine how the worldline of a propagating light ray is situated with respect to the spatial and temporal dimensions. If the speed of light is denoted by c, the equation for the propagation of a light ray along the x-axis is $x = ct$. This equation shows that there are two possibilities: on the time axis we can plot either t or ct. In order to obtain uniformity of all four dimensions of spacetime in the sense that all are measured in the same units – length – the researchers choose to plot ct on the time axis. With this choice the worldline of a light ray will form an angle of 45° with the time axis, as the equation $x = ct$ shows.

Figure 4.4 depicts the expansion of a light sphere like the one used in the thought experiment designed by the traditional research team to detect absolute uniform motion, as discussed in Chap. 3. At the moment $t = 0$ s a light signal is emitted and a spherical light wave starts to expand. At different moments of time we perceive light spheres with

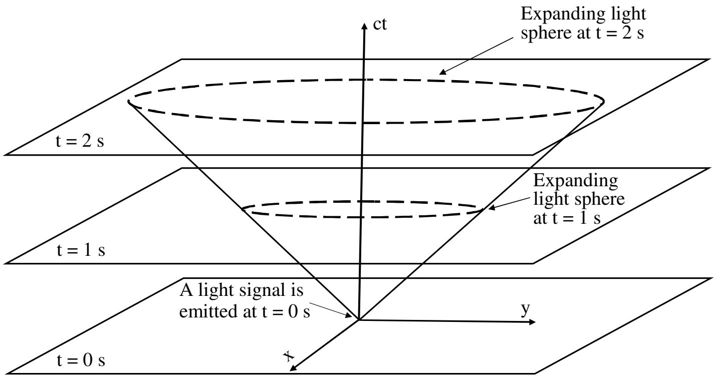

ct
Expanding light sphere at t = 2 s
t = 2 s
Expanding light sphere at t = 1 s
t = 1 s
A light signal is emitted at t = 0 s
y
x
t = 0 s

Fig. 4.4. The perceived expanding light sphere is in fact a four-dimensional light 'cone' in spacetime

different radii. However, there are no expanding three-dimensional light spheres in spacetime. The whole history of the emitted light signal is entirely realized as a four-dimensional light 'cone' there. Our instantaneous three-dimensional spaces corresponding to different moments of time 'cut' different cross-sections from the light cone, which we interpret as the expanding light sphere at different moments.

The radical research team immediately realizes that if the four-dimensional light cone is not just an exercise in abstract thinking but is part of the true reality, then the apparent paradox depicted in Fig. 3.6 that the A- and B-observers have different light spheres (which originated from the same light signal) finds a natural explanation. As shown in Fig. 4.5 the non-coinciding three-dimensional spaces of A and B 'cut' different cross-sections (different three-dimensional light spheres) from the four-dimensional light cone.[6] So, what is the same light signal is the light cone, but the three-dimensional spaces of different observers in relative motion will 'cut' it at different angles and will therefore have different three-dimensional light spheres.

The researchers of the radical team themselves are not certain whether such an extreme idea as the concept of spacetime will turn out to have anything to do with the external world, but their aim is to see to what experimental predictions the exploration of the internal logic of this idea will lead them. And here again they have one more oppor-

[6] In Fig. 4.5 the two cross-sections are not of the same shape – one is a circle, whereas the other is an ellipse. In the pseudo-Euclidean geometry of spacetime, however, two three-dimensional spaces intersect the light cone in two light *spheres*.

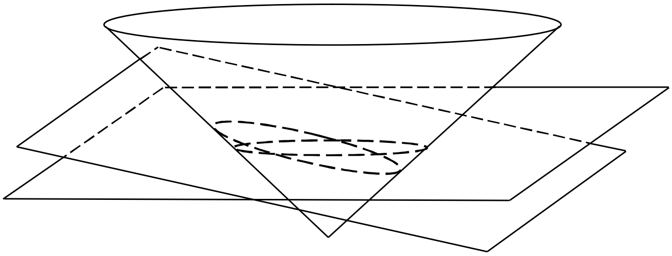


Fig. 4.5. The instantaneous three-dimensional spaces of two observers in relative motion intersect the four-dimensional light cone of a light signal at different angles. The two cross-sections are interpreted by the observers as two light spheres

tunity to test its internal consistency – the paradox of the two light spheres in Fig. 3.6 has found an elegant explanation. The researchers from this team are certainly aware that internal consistency does not prove a hypothesis, but they are also aware that any hypothesis that has any chance of being experimentally confirmed must be internally consistent.

So far, the concept of light cone only reflects the future history of a light signal emitted from a given point. But what about the histories of all light signals which arrive at the same point? And that is an important question since what we see at a given moment are all light signals which reach our eyes at that moment. What we see in Fig. 4.6 is the light cone which is associated with a given point of spacetime, say the event O. It consists of the past light cone, which contains the histories of all light signals arriving at O, and the future light cone, which contains the whole history in time of a light signal emitted at

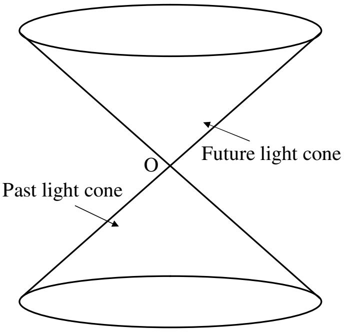


Fig. 4.6. A light cone is associated with an event of spacetime

O. If the event O is considered to be the present moment, then the past light cone contains the past histories of the light signals reaching O, whereas the future light cone consists of the future history of the light signal emitted at O.

The light cone clearly demonstrates something that was anticipated by the radical research team – that what we 'perceive' is not one space but consists of points of spaces belonging to different moments of time (as shown in Fig. 3.2). As Fig. 4.6 shows, an observer at event O will see the past light cone. So what the observer sees at O is what we all see at every instant of our everyday experience – all objects around us and the three-dimensional space. However, the past light cone does not constitute a space since the three-dimensional space is defined as all points that exist at a *single* moment of time, whereas the events comprising the light cone correspond to different moments of time. Stated another way, at O the observer will be convinced that he perceives a three-dimensional world, but this is not so since a three-dimensional world is defined at a single moment of time. Therefore, what we perceive is not a three-dimensional space (or a three-dimensional world), but a special region of spacetime – the past light cone (Fig. 4.7). This fact further undermines the intuitive argument that the world is obviously three-dimensional since it is what we see.

The light cone is an element of spacetime and does not depend on the introduction of reference frames which allow us to describe the timelessly existing spacetime in our ordinary three-dimensional language. So, a light cone is associated with an event, not with a worldline or reference frame.

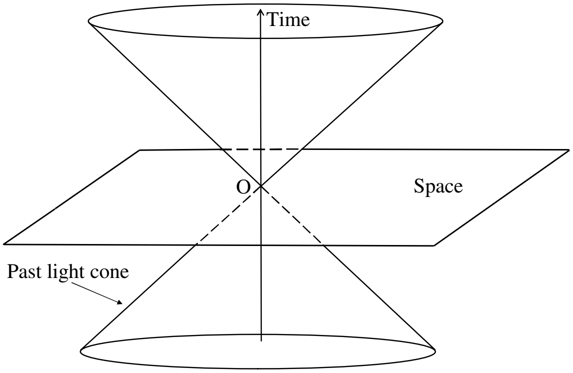


Fig. 4.7. An observer at event O sees the past light cone, not a three-dimensional world (or a three-dimensional space)

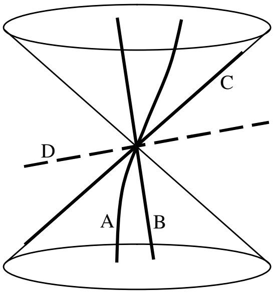


Fig. 4.8. Three different kinds of worldline

The very fact that, unlike the speeds of any other particles or disturbances, the speed of light is the same in all inertial reference frames makes the researchers of the radical team suspect that perhaps there is more physics behind this fact. That is why they look at how different worldlines, believed to contain the whole histories in time of different particles, compare to the worldlines of light rays. (One can think of the light cone as consisting of an infinite number of worldlines of light rays.)

Looking at Fig. 4.8, it is easy to realize that, in terms of their relation to the worldline of a light ray, one can distinguish three kinds of worldline. Worldlines A and B lie in the light cone and represent particles whose speeds are smaller than the speed of light. Worldline C lies on the light cone and corresponds either to a light ray or to a particle that moves at the speed of light. Worldline D is situated outside of the light cone and, if every worldline represents a particle, should correspond to a particle whose speed is greater than the speed of light.

The scientists of the radical research team now face perhaps the most difficult task in their analysis of the features of spacetime. They have to start to build the mathematical formalism of a four-dimensional space with three spatial dimensions and one temporal dimension. In all figures they have used so far the temporal and the spatial dimensions looked exactly the same. On the one hand, this should be so, since all dimensions of a space must be given at once. In the case of spacetime all dimensions must also exist equally; otherwise, if the time dimension were not entirely given, spacetime would be three-dimensional, not four-dimensional. On the other hand, however, it is taken as self-evident that the temporal dimension is different from the spatial dimensions. But when it comes to explaining in what sense time is different, the situation becomes reminiscent of Saint Augustine's confusion over time [50]: "If no one asks me, I know; if I wish to explain to him who asks, I know not." Most people believe that this difference is evident from the fact that we perceive the time dimension moment after moment, whereas we 'see' the spatial dimensions at once. This belief, however, cannot be used as an argument since what we 'see', as shown in Fig. 4.7, is not a space but the past light cone.

The distinction between the temporal dimension and the spatial dimensions is best demonstrated by the lack of complete freedom in choosing a time direction. In Euclidean space we are completely free to choose, say, the direction of the y-axis. This is not the case in spacetime. Assume that we decide to introduce an inertial reference frame whose time axis is along worldline D lying outside of the light cone (see Fig. 4.8), which means that the worldline B will lie entirely in our instantaneous three-dimensional space. But since worldline B contains the whole history of its corresponding particle, such a choice of the time direction would mean that the particle existing at all moments of its history would momentarily (and therefore *simultaneously*) appear in our space coming from nowhere and then would disappear again. The lack of any evidence for such occurrences suggests that we cannot choose a time direction along a worldline lying *outside* of the light cone. By the same argument, however, we should assume that the existence of particles moving at speeds greater than the speed of light is highly unlikely. If we choose the time axis of our inertial reference frame along worldline B, worldline D, containing the entire history of a hypothetical superluminal particle, will lie entirely in our instantaneous three-dimensional space. This would mean that at a given instant of our time the superluminal particle will appear in our instantaneous space *simultaneously* existing at all moments of its history. Such a thing, involving ordinary macroscopic objects, has never been observed. If we can reach the same conclusion in the case of microscopic particles, then we will have good reason to believe that superluminal particles do not exist. And we will also have a good question to ask: Why is the existence of a whole class of possible worldlines in spacetime not possible?

The next question the radical research team asks is whether we can choose a time axis along a worldline lying on the light cone. At this stage of their analysis an answer does not seem easily deducible. However, there are clear indications that such a choice will not be possible either. An inertial reference frame associated with the worldline of a light ray will be moving at the constant speed c relative to all other inertial reference frames and therefore will be in this sense a privileged reference frame. What is even more serious is that the association of an inertial reference frame with a light ray leads to a contradiction with the first consequences from the 'no-absolute-motion' interpretation of Galileo's principle of relativity – that the speed of light is constant in all inertial reference frames. As a light ray by definition would be at rest in an inertial reference frame associated with it, it follows that light would not be moving in an inertial reference frame whose time axis is along the worldline of a light ray.

Therefore the radical research team arrives at the conclusion that all possible time directions lie inside the light cone. This suggests that all worldlines inside the light cone should be regarded as objectively different from the other two types of worldline. Therefore, the constant speed of light does appear to play some important role in nature since it is the light cone that defines the different kinds of worldline. Let us call the worldlines lying inside the light cone *time-like* and the worldlines on the light cone itself – *light-like*. The hypothetical worldlines situated outside of the light cone can be called *space-like* worldlines.

Having analyzed the basic objects of spacetime and concluded that its temporal and spatial dimensions are in one respect equal (they are all given at once) but in another respect different (the time dimension cannot be interchanged with a spatial dimension),[7] the radical research team can start to work on the mathematical description of spacetime.

4.2 Derivation of the Lorentz Transformations

The researchers' idea for a mathematical description of spacetime is quite simple. They want to see whether the assumption that the external world is four-dimensional leads to testable predictions. And as such a world, which we called spacetime, is given at once like the two-dimensional surface of the page you are now reading, the first thing that can be done is to study its *geometry* – the relations between the worldlines embedded in spacetime. Since there is an obvious similarity between lines in the ordinary Euclidean space and worldlines in spacetime, the researchers will start with the well-known relations between lines drawn on a two-dimensional Euclidean space (surface) and examine the corresponding relations between worldlines in two-dimensional spacetime (where two spatial dimensions of the four-dimensional spacetime have been suppressed). Then they will translate the established

[7] The temporal and spatial dimensions of spacetime should be treated as Taylor and Wheeler put it [21, p. 18]: "Equal footing, yes; same nature, no."

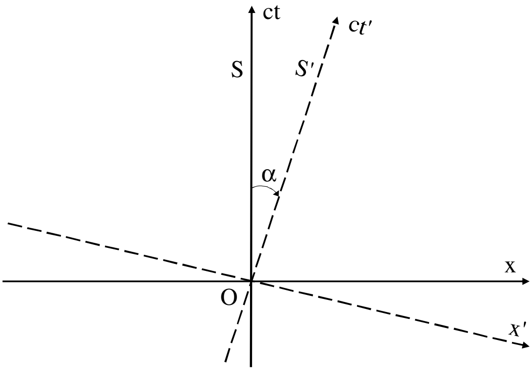


Fig. 4.9. The time axes of two observers in relative motion are chosen along the observers' worldlines (which coincide with the time axes in the figure)

relations between the worldlines into the ordinary three-dimensional language in order to see whether the spacetime hypothesis makes observable predictions.

Consider two parallel worldlines. They correspond to two observers who are at rest with respect to each other. If the worldlines form an angle, the observers are in relative motion. When we introduce inertial reference frames S and S' with time axes along the two worldlines, it is easily seen that the two reference frames are rotated with respect to each other as shown in Fig. 4.9.

This suggests to the radical research team to try to find the transformations between two inertial reference frames in relative motion by using the known transformation between two coordinate systems K and K' which are rotated with respect to each other. Such transformations are crucial for any theory that claims to describe objective facts. The length between two points in a two-dimensional Euclidean space, for instance, is an objective fact. And since we can choose different coordinate systems to calculate the length, the calculations should obviously not depend on our choice of coordinate system and should give the same length. In other words, the length should remain invariant when we choose a different coordinate system to calculate the same length.

The most general transformations between two coordinate systems involve translation and rotation of the systems. However, the radical research team would like first to examine the transformations of rotation, since a rotation in spacetime corresponds to relative motion between two inertial reference frames.

Consider a coordinate system K' which is rotated through an angle α with respect to another coordinate system K (Fig. 4.10). A point P has coordinates (x, y) in K and (x', y') in K'.

The transformations of the coordinates in K into the coordinates in K', i.e., $K \rightarrow K'$, can be deduced directly from Fig. 4.10:

$$\begin{cases} x' = x\cos\alpha - y\sin\alpha \ , \\ y' = x\sin\alpha + y\cos\alpha \ . \end{cases} \tag{4.1}$$

It is easily checked that these transformations leave the distance between two points in the two-dimensional Euclidean space invariant. The length OP in K' is $l'^2 = x'^2 + y'^2$. For the length of OP in K, the transformation (4.1) gives $l^2 = x^2 + y^2$.

Now the radical research team wants to see whether the transformations (4.1) in Euclidean space can be directly used in spacetime for the case of two inertial reference frames S and S' which are rotated with respect to each other as shown in Fig. 4.9. However, simply replacing y with ct in (4.1) will not work since such a substitution still preserves our freedom to interchange the two axes x and ct. Nothing changes when x and y (representing two spatial dimensions) are interchanged, but the same thing cannot be done in spacetime since ct and x represent dimensions of different nature. At this point the radical research

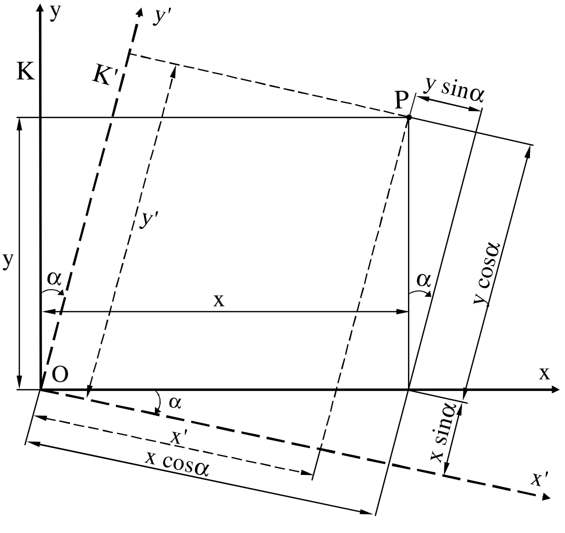


Fig. 4.10. Two coordinate systems are rotated in a two-dimensional Euclidean space

team faces a serious difficulty. The problem they have to resolve has not been encountered by anyone before them. The closest situation is the one when the issue with real and imaginary numbers was dealt with. The geometric representation of complex numbers involves an axis for the real numbers and another one for the imaginary numbers; the two axes cannot be interchanged. Although the two cases are quite different, the radical research team decides to try the imaginary unit $\mathrm{i} = \sqrt{-1}$ to distinguish the temporal and the spatial dimensions in spacetime.

It does not matter whether we choose the coordinates $(\mathrm{i}ct, x, y, z)$ or $(ct, \mathrm{i}x, \mathrm{i}y, \mathrm{i}z)$; what is important is that the temporal and the spatial coordinates cannot be interchanged, which reflects the different nature of the temporal and the spatial dimensions of spacetime. However, the multiplication of a coordinate by i makes it *imaginary*, whereas what we measure, including temporal and spatial intervals, is represented only by *real* numbers. The radical research team knows that they need the imaginary unit i only as a translation factor to distinguish between the temporal and the spatial coordinates when they translate Euclidean into spacetime expressions. That is why they decide to use only the expression (containing real quantities), which multiplies i, after the translation has been completed.

The traditional research team will be outraged by such an abuse of mathematics. They will even point out to their colleagues of the radical research team that, for the sake of their common desire to understand the physical meaning of Galileo's principle of relativity, they have been following the exotic hypotheses of their radical colleagues, but with their artistic approach toward mathematics the radical team has gone too far. The traditional scientists would argue that mathematics has firm rules that cannot be twisted to accommodate bizarre hypotheses which have nothing to do with the real world. The most probable reaction of the radical research team might be to explain patiently that what matters the most in mathematics is internal consistency; so, if something is introduced in a self-consistent manner and predictions are deduced, it will be wise to let experiment have its say about both the initial hypothesis and its mathematical formulation. One day both teams would witness cases when scientists introduced even newer mathematical objects, initially encountering the opposition of mathematicians but later being accepted as respected members of the growing mathematical family. But for now the radical research team will carry on with their goal to test experimentally the hypothesis that the four-dimensional spacetime represents the true world.

The standard rotation transformations (4.1) are not in the best form for the translation into the spacetime rotation transformations. The reason is that they contain $\sin\alpha$ and $\cos\alpha$, whereas for the spacetime transformations (4.1) should contain $\tan\alpha = \Delta x/\Delta y$, since the translation of $\tan\alpha$ will include the relative velocity between the inertial reference frames S and S' (Fig. 4.9); that this is so is clear from the fact that $\tan\alpha$ will contain the ratio $\Delta x/c\Delta t$, which is v/c, as can be seen in Fig. 4.9. And we have already seen that the angle between the worldlines of two particles corresponds to the relative velocity of the particles.

The rotation transformations are easily rewritten in terms of $\tan\alpha$. By making use of the trigonometric identity $\sin^2\alpha + \cos^2\alpha = 1$, the first equation of (4.1) becomes:

$$\begin{aligned} x' = x\cos\alpha - y\sin\alpha &= \frac{x\cos\alpha - y\sin\alpha}{(1)^{1/2}} \\ &= \frac{x\cos\alpha - y\sin\alpha}{(\cos^2\alpha + \sin^2\alpha)^{1/2}} = \frac{\cos\alpha(x - y\tan\alpha)}{\cos\alpha(1+\tan^2\alpha)^{1/2}} \\ &= \frac{x - y\tan\alpha}{(1+\tan^2\alpha)^{1/2}} \ . \end{aligned}$$

As the second equation of (4.1) can be rewritten in the same way, we obtain the new form of the rotation transformations:

$$x' = \frac{x - y\tan\alpha}{(1+\tan^2\alpha)^{1/2}} \ , \qquad (4.2)$$

$$y' = \frac{y + x\tan\alpha}{(1+\tan^2\alpha)^{1/2}} \ . \qquad (4.3)$$

Now the radical research team can translate the Euclidean rotation transformations (4.2) and (4.3) into spacetime rotation transformations by one of the two substitutions:

$$x \longrightarrow \mathrm{i}x \ , \qquad y \longrightarrow ct \ , \qquad (4.4)$$

or

$$x \longrightarrow x \ , \qquad y \longrightarrow \mathrm{i}ct \ . \qquad (4.5)$$

Let us now use the first substitution (4.4). As we expected, the angle between the two worldlines, along which the time axes of the inertial reference frames S and S' are chosen (Fig. 4.9), is related to the relative velocity v of the particles that correspond to the two worldlines:

$$\tan\alpha = \frac{\Delta x}{\Delta y} \to \frac{\mathrm{i}\Delta x}{c\Delta t} = \mathrm{i}\frac{v}{c} = \mathrm{i}\beta\ , \tag{4.6}$$

where we have used the now common notation $\beta = v/c$. We can now translate the equation (4.2) for the transformation of the x coordinate:

$$\mathrm{i}x' = \frac{\mathrm{i}x - ct(\mathrm{i}\beta)}{[1+(\mathrm{i}\beta)^2]^{1/2}} = \frac{\mathrm{i}(x-\beta ct)}{(1-\beta^2)^{1/2}}$$

or

$$\mathrm{i}\left[x' = \frac{x-\beta ct}{(1-\beta^2)^{1/2}}\right]\ .$$

For the equation (4.3) transforming the y coordinate, we obtain:

$$ct' = \frac{ct + \mathrm{i}x(\mathrm{i}\beta)}{[1+(\mathrm{i}\beta)^2]^{1/2}} = c\frac{t-\dfrac{\beta}{c}x}{(1-\beta^2)^{1/2}}\ .$$

The translation of the rotation transformations $K \to K'$ in the two-dimensional Euclidean space gives the rotation transformations $S \to S'$ in a two-dimensional spacetime:

$$x' = \frac{x-\beta ct}{(1-\beta^2)^{1/2}}\ , \tag{4.7}$$

$$t' = \frac{t-\dfrac{\beta}{c}x}{(1-\beta^2)^{1/2}}\ , \tag{4.8}$$

where we have applied the rule that we will consider only the expression multiplying the imaginary unit i.

We immediately recognize that (4.7) and (4.8) are the Lorentz transformations. We have obtained them through the substitution $x \to \mathrm{i}x$ and $y \to ct$ (4.4), but it is easily seen that the other substitution (4.5) also leads to the Lorentz transformations (4.7) and (4.8). The radical research team is especially delighted at the fact that both substitutions (4.4) and (4.5) translate the Euclidean rotation transformations (4.2) and (4.3) into the *same* spacetime rotation transformations (4.7) and (4.8). As the researchers expected, it does not matter whether the time or the spatial coordinates are multiplied by the imaginary unit i; the purpose is to find a way to distinguish between the temporal and the spatial dimensions. So far the consistency of their approach has been confirmed.

An inspection of the Lorentz transformations demonstrates that the constant c which coincides with the constant speed of light does play a fundamental role in spacetime, since it is present in the transformations between two inertial reference frames in relative motion. The traditional research team will not miss the opportunity to point out the obvious reason why c participates in the Lorentz transformations – it was the radical research team who defined the time axis in terms of ct and later included c in the two substitutions (4.4) and (4.5). Hence there is nothing exciting about the presence of c in (4.7) and (4.8). The radical research team will again patiently explain that what they find exciting is that the speed c represents some *limiting speed* which holds for the motion of any objects and processes – for speeds greater than c, the Lorentz transformations break down. It is true that c was initially introduced when the time axis was represented as ct, but its role as a limiting speed was not expected (and not presupposed).

The radical research ream realizes that the existence of a limiting velocity has implications not only for physics itself but also for our world view. Such a limiting velocity leads to a new (relativistic) division of events into past, present, and future. All events on and in the past light cone can influence event O and are therefore past events with respect to O, which is considered as present. Event O can influence all events on and in the future light cone, which means that the events on and in the future light cone are future events with respect to O. The events occupying the region outside of the light cone are

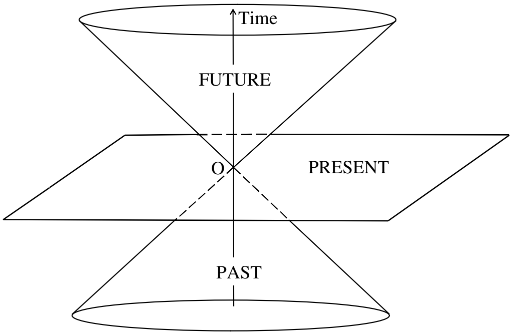


Fig. 4.11. The fact that c is a limiting speed implies a new division of events – the past light cone contains all past events with respect to event O, whereas all events lying on and inside the future light cone are future events with respect to O. Only the event O is regarded as present. The events outside of the light cone are neither past nor future. This new division of events shows that the presentist view is in trouble because the present, or the three-dimensional world at the moment ‘now’, is defined in terms of the pre-relativistic division of events into past, present, and future

neither past nor future. One may argue that they can be regarded as a relativistic version of the present events. However, if the region outside of the light cone is considered to be the relativistic present, then it is not a three-dimensional world, but a four-dimensional region of spacetime, as shown in Fig. 4.11. This situation shows that the presentist view, according to which it is only the present – the three-dimensional world existing at the constantly changing moment 'now' – that exists, is not consistent with the existence of a limiting velocity. It should be stressed that a three-dimensional world is defined only in terms of the pre-relativistic division of events into past, present and future, as shown in Fig. 4.11.

Another point that should be emphasized is that, unlike the pre-relativistic division of events, the relativistic division does not affect the existence of the events – the events in the past light cone, the event O, and the events in the future light cone are all *equally* existent. This can be seen in Fig. 4.12. Consider two light cones whose present events are O and O', respectively. The equal status of existence of the events lying in the different regions of the left-hand light cone is demonstrated by the fact that parts of its past and future light cone as well as part of its outside region and the event O all lie in the region that is outside of the right-hand light cone.

Once the radical research team have obtained the Lorentz transformations, they can derive all consequences of special relativity in the way it is done in the standard books on relativity. However, this is not their goal. As discussed earlier they want to translate [by making use of the substitution (4.4) or (4.5)] relations between lines in the ordinary Euclidean space into relations between worldlines in spacetime. Then they will express the relations between the worldlines of

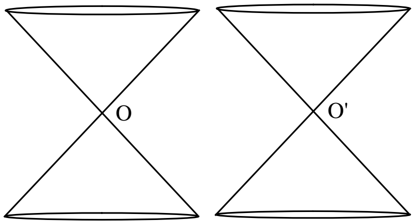


Fig. 4.12. The events of spacetime lying in the past and future light cone are not objectively divided into past, present (event O), and future. This becomes evident if we consider a second light cone whose present event is the event O'. Parts of the past and future light cone, as well as part of the area outside of the first light cone, lie outside of the second light cone

physical objects in terms of the ordinary three-dimensional language which will allow them to test those relations experimentally. In this way they will be able to test the hypothesis that the worldlines are real four-dimensional objects belonging to a real four-dimensional world.

4.3 Four-Dimensional Distance and Three Kinds of Length

Before starting to examine the relations between lines in Euclidean space and their translations into relations between worldlines in spacetime, let us first translate the expression for the length between two points (which is an invariant in Euclidean space) into the corresponding expression for the four-dimensional length between two events in spacetime (which should be an invariant in spacetime). By making use of the first substitution (4.4), we obtain

$$l^2 = x^2 + y^2 \longrightarrow (\mathrm{i}x)^2 + (ct)^2 = -x^2 + c^2t^2 \; .$$

So the translation of the square of the two-dimensional Euclidean distance l^2 gives the square of the distance s^2 in two-dimensional spacetime:

$$s^2 = c^2t^2 - x^2 \; .$$

The expression for the distance in a four-dimensional spacetime, also called the interval, is

$$s^2 = c^2t^2 - x^2 - y^2 - z^2 \; . \tag{4.9}$$

By using the second substitution (4.5), we obtain a different expression for the length in a two-dimensional spacetime:

$$s^2 = -c^2t^2 + x^2 \; ,$$

whose four-dimensional generalization is

$$s^2 = -c^2t^2 + x^2 + y^2 + z^2 \; . \tag{4.10}$$

The fact that the radical research team obtained two different expressions for the four-dimensional distance (4.9) and (4.10) might not come as a surprise to them. What they wanted to achieve was to have a mathematical description which distinguished between the temporal and the spatial dimensions. And indeed the only difference between the two expressions is the different signs in front of the temporal and the spatial coordinates. The expressions (4.9) and (4.10) are equivalent, and today, books on relativity use either (4.9) or (4.10). The signs in the expression for the length are called the signature of a space. The signature of Euclidean space (for the length $l^2 = +x^2 + y^2 + z^2$) is $(+ + +)$, whereas the signature of spacetime is either $(+---)$ or $(-+++)$.

In order to check that the interval (4.9) [or (4.10)] is invariant under the Lorentz transformations, we can generalize (4.7) and (4.8) by including the y and z coordinates. When two inertial reference frames are in relative motion along their x axes, the Lorentz transformations are

$$\begin{aligned} x' &= \frac{x - \beta ct}{(1-\beta^2)^{1/2}} \, , \\ t' &= \frac{t - \dfrac{\beta}{c}x}{(1-\beta^2)^{1/2}} \, , \\ y' &= y \, , \\ z' &= z \, . \end{aligned} \tag{4.11}$$

The interval in S' is $s'^2 = c^2t'^2 - x'^2 - y'^2 - z'^2$. It is easy to check that using (4.11) to transform it into the interval in S gives (4.9).

As the expression for the distance contains important information about the geometry of space,[8] let us examine more closely the difference between the length

$$l^2 = x^2 + y^2 + z^2$$

in Euclidean space and

$$s^2 = c^2t^2 - x^2 - y^2 - z^2$$

in spacetime. The signature of the two spaces is different and this has significant implications. As seen from the expression for l in Euclidean space, the distance l between two points is zero if and only if the points coincide – the x, y, and z components of the length l must all be zero in order that l be zero. In spacetime, however, this is not the case. Consider a time-like, a light-like, and a hypothetical space-like worldline all passing through the tip of the light cone associated with the event O (Fig. 4.13).

Let us calculate the length along each of the worldlines. For the spacetime length between events O and A lying on the time-like worldline, we have

[8] For instance, the length $l^2 = x^2 + y^2$ between points O and P in Fig. 4.10 is in fact the Pythagorean theorem.

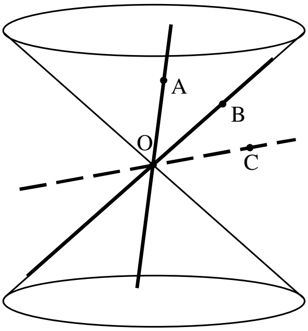


Fig. 4.13. Three kinds of length in spacetime

$$\Delta s^2 = c^2 \Delta t^2 - \Delta x^2 > 0 \ . \tag{4.12}$$

That $\Delta s^2 > 0$ can most easily be seen if we choose an inertial reference frame whose time axis is along the time-like distance OA. Then $\Delta x = 0$, since the events O and A occur at the origin of the reference frame ($x_O = x_A = 0$) and obviously $\Delta s^2 = c^2 \Delta t^2 > 0$. In any other inertial reference frame, the particle represented by the time-like worldline will move at speed $v < c$, which means that the time period Δt for which the particle travels the distance Δx when multiplied by c gives a length $c\Delta t$ which is greater than Δx. Therefore, for any choice of reference frame, the spacetime interval along a time-like worldline is $\Delta s^2 > 0$.

For the spacetime length between events O and B lying on the light-like worldline, we will obviously have $\Delta s^2 = 0$. In all inertial reference frames the speed of light is constant and the distance Δx travelled by a light ray, represented by a light-like worldline, is $c\Delta t$. The spacetime length between events O and C, which are on the hypothetical space-like worldline, is negative: $\Delta s^2 < 0$. The reason is that in any inertial reference frame the hypothetical particle represented by the space-like worldline will move faster than light and the distance Δx covered in time Δt will always be greater than the distance $c\Delta t$ which a light signal travels for the same time Δt. This can be most clearly seen in the inertial reference frame whose three-dimensional space instantaneously contains a hypothetical space-like worldline. In such a case $\Delta t = 0$ since $t_O = t_C$ and $\Delta s^2 = -\Delta x^2 < 0$.

The fact that the length along the three kinds of worldline is different confirms the conclusion that there exists an objective difference between these worldlines. We have not gained any insight into whether or not space-like worldlines (and the corresponding superluminal particles) exist. As the concept of spacetime itself does not forbid their existence it will once again be experiment that decides. A frequent argument against superluminal particles and therefore against the existence of space-like worldlines is the so-called relativistic causality. As shown in Fig. 4.13, in all reference frames event O occurs *before* events A and B and this is interpreted to mean that O *causes* A and B. When two events lie on a hypothetical space-like worldline (or are space-like separated), however, which occurs first becomes frame-dependent – in some reference frames O will occur before C, whereas in other frames it will be C that occurs before O. This clearly violates the accepted belief that a cause must precede its effects. The violation of the cause–effect assumption and the fact that c is a limiting speed in the Lorentz transformations (4.11) are the reasons for the rejection of motions faster than that of light (which means a rejection of space-like worldlines as well).

For the researchers of the radical team the cause–effect issue looks different. In terms of spacetime all events of a worldline (not only space-like) are given at once and do not objectively occur one by one. In this sense an event of a worldline is not caused by another event of the worldline (just as a point on a line is not caused by another point on the same line). If spacetime is real, events do not occur – they are all there in spacetime; they 'occur' only when a reference frame is introduced in order to *describe* the timeless existence of all events in the three-dimensional language of our perceptions. However, as discussed above, the radical research team believes that it is unlikely that space-like worldtubes of macroscopic objects exist – the existing macroscopic experimental evidence overwhelmingly supports the cause–effect assumption and rejects the existence of space-like worldtubes. As the researchers from this team are interested in seeing how the geometrical structure of spacetime itself excludes the class of space-like worldlines, they would like to find out what experiment tells us about superluminal motions of microscopic particles. No matter what the verdict may be, there will be tough questions to answer. If there are no space-like worldlines, what in the geometrical structure of spacetime disallows those worldlines? If space-like worldlines of microscopic particles turn out to exist, then a superluminal version of the Lorentz transformations should be derived. In this case the question will be: What prevents the existence of space-like worldlines at the macroscopic level?

The fact that there are three kinds of length in spacetime reveals the difference between spacetime geometry and Euclidean geometry. Although the radical research team anticipated such a difference due

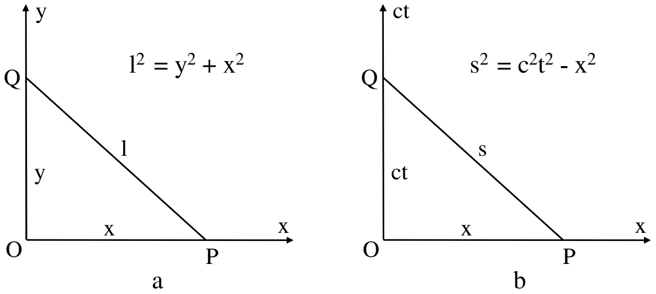


Fig. 4.14. The Pythagorean theorem in Euclidean space (**a**) and in spacetime (**b**)

to the different nature of the temporal and the spatial dimensions, they found the spacetime geometry quite counter-intuitive. The difference between time-like and space-like lengths was expected, but the light-like length came as a complete surprise. It is simply impossible to visualize how the distance between two *non-coinciding* points on a light-like worldline (say O and B in Fig. 4.13) can be zero no matter how far apart the points may seem to be. This fact shows the radical research team that they have to be very careful with their spacetime diagrams. Any conclusions deduced from such a diagram should be obtained independently by making use of the Lorentz transformations. The light cone drawn on the Euclidean surface of a piece of paper provides only an idea of what the light cone may look like in the different (let us call it *pseudo-Euclidean*) geometry of spacetime. The same is true for all other spacetime diagrams we will discuss in this book.

The realization that the length between two events on a light-like worldline is always zero provides an additional justification for the conclusion that no reference frame can be associated with a light signal. If we were to attach an inertial reference frame to a moving light ray, we should choose the time axis of the reference frame along the light-like worldline of the light ray. This would mean that the spacetime interval between the events O and B (Fig. 4.13) as determined in that reference frame would be $\Delta s^2 = c^2\Delta t^2$, since $\Delta x = 0$ there. But the spacetime interval in this case is zero in any other reference frame and due to its invariance it should be zero in the reference frame associated with the light ray as well. Therefore, in this reference frame $\Delta t^2 = 0$, which means that no time can be defined there. For this reason no reference frame can be associated with a light-like worldline.

In the pseudo-Euclidean geometry of spacetime, the Pythagorean theorem also looks different. In the ordinary Euclidean geometry, the square of the hypothenuse of a triangle (the length between points P and Q in Fig. 4.14a) is equal to the sum of the squares of the two other sides of the triangle: $l^2 = x^2+y^2$. When we translate this expression through the substitution (4.4), we obtain the expression for the Pythagorean theorem in the pseudo-Euclidean geometry of spacetime: $s^2 = c^2t^2 - x^2$ (Fig. 4.14b).

The pseudo-Euclidean nature of the spacetime geometry and the difficulty in representing the relations between worldlines in spacetime on a Euclidean surface is also evident from an examination of the Lorentz transformations (4.7) and (4.8) [22, p. 50]. When the time axis of reference frame S' is rotated with respect to the time axis of frame S, the x' axis rotates in the opposite direction. What is important, however, is that relative motion of two particles in the ordinary space is represented by the particle worldlines, which are *rotated* with respect to each other in spacetime.

Since it is helpful to visualize the pseudo-Euclidean relations between worldlines in spacetime, we will try to depict them on the Euclidean surface of the page in such a way that their main results are preserved. We will use the following rule which applies only to two reference frames and which can be regarded as a mnemonic device. In Fig. 4.14a the two smaller sides of the triangle OPQ are along the x and y axes. As x and y are *orthogonal*, the calculation of the length $l^2 = x^2+y^2$ is obtained through the Pythagorean theorem. The pseudo-Euclidean version of the Pythagorean theorem is also used in the calculation of the spacetime interval s, which is the distance between events P and Q. Let the spacetime distance $s = PQ$ be calculated in two inertial reference frames S and S' in relative motion. Due to the invariance of the spacetime interval $s^2 = s'^2$, we can write

$$c^2t^2 - x^2 = c^2t'^2 - x'^2 \ ,$$

which can be rewritten as

$$c^2t^2 + x'^2 = c^2t'^2 + x^2 \ . \tag{4.13}$$

Both the left-hand and the right-hand sides of (4.13) look like the right-hand side of the Pythagorean theorem $l^2 = x^2 + y^2$. As x and y are orthogonal, we can draw the x' axis perpendicularly to ct and x – perpendicularly to ct' (Fig. 4.15).

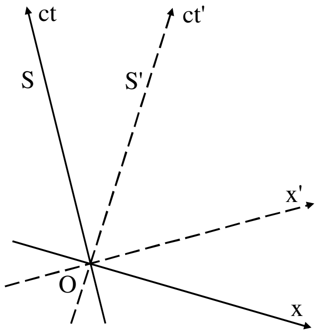


Fig. 4.15. The x axis of S is perpendicular to ct', whereas the x' axis of S' is perpendicular to ct

This way of drawing the axes of two inertial reference frames in relative motion on the Euclidean surface of the page correctly represents the relativity of simultaneity and, as we will see in the following sections, the time dilation and the length contraction effects. That Fig. 4.15 adequately depicts the relativity of simultaneity can be demonstrated in the case of the thought experiment involving two spacecraft A and B discussed in Chap. 3. For the B-observers, the propagating light sphere, emitted when the middle points of A and B instantaneously coincide, simultaneously reaches the rear and front end points of B. In other words, the events BR (light hits the rear end point of B) and BF (light arrives at the front end point) are simultaneous in spacecraft B, as shown in Fig. 3.7. For the A-observers, however, the expanding light sphere reached the rear end point of B first, and therefore the event BR happened *before* the event BF. As shown in Fig. 4.16, the way we decided to draw the temporal and spatial coordinates of two inertial reference frames in relative motion adequately reflects the way the observers in the two frames determine the order in which events fall in their instantaneous spaces. If we drew the x and x' axes orthogonal to the t and t' axes, respectively, the order of events for the A-observers would be wrong.

4.4 Y 'Dilation' in Euclidean Space and Time Dilation in Spacetime

The idea of the radical research team is to translate relations between lines in two-dimensional Euclidean space into the corresponding relations between worldlines in spacetime. By doing this they hope to

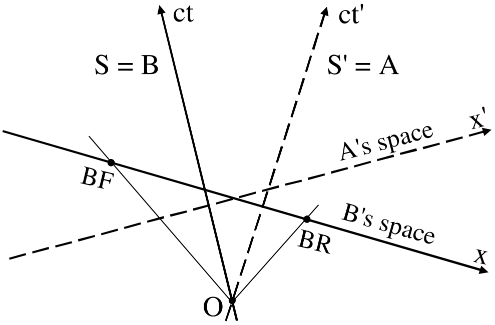


Fig. 4.16. At event O, when the middle points of spacecraft A and B coincide, a light signal is emitted. For the B-observer the expanding light sphere *simultaneously* reaches the rear end point of B (event RB) and the front end point of B (event FB). The A-observer, however, determines that event BR occurs before event BF

obtain experimental predictions that can be used to test the reality of the worldlines and hence the reality of spacetime. For this purpose we will first derive the relations between lines in two-dimensional Euclidean space and then translate them to find the corresponding relation between worldlines in spacetime. For all translations we will use the substitution (4.4).

Consider two coordinate systems K and K' rotated with respect to each other by an angle α, as shown in Fig. 4.17. An observer in K measures the length OA of a rod lying along the y-axis of his coordinate system and finds that it is y. In this case the length of the rod coincides with its y component since the rod's length does not have an x component in K. An observer in the coordinate system K' finds that the length of the rod has both x- and y components in K'.

The K'-observer wants to compare the y components of the rod in K and K'. In order to do that he can use the rotation transformation $K \rightarrow K'$ of the y components of K and K' (4.3) which *projects* point A onto point A' on the y'-axis. Noting that the rod does not have an x component, the K$'$-observer obtains

$$y' = \frac{y}{(1 + \tan^2 \alpha)^{1/2}} \, . \tag{4.14}$$

The relation (4.14) can also be obtained by noting how the K'-observer will determine that relation. He can move a line parallel to his x'-axis upwards until it reaches point A. Then the K'-observer determines where the line parallel to x' intersects the y'-axis and he finds that the y

component of the rod's length in K' is $OA' = y'$. As shown in Fig. 4.17, $y' = y\cos\alpha$. Using the trigonometric equality $\cos^2\alpha + \sin^2\alpha = 1$ again, we can write $\cos\alpha$ in the following way:

$$\cos\alpha = \frac{\cos\alpha}{(1)^{1/2}} = \frac{\cos\alpha}{(\cos^2\alpha + \sin^2\alpha)^{1/2}} = \frac{1}{(1+\tan^2\alpha)^{1/2}} \ .$$

We then obtain

$$y' = \frac{y}{(1+\tan^2\alpha)^{1/2}} \ ,$$

which, as expected, coincides with (4.14). The relation (4.14) between the y components of the rod in K and K' can be calculated using the invariance of the distance in Euclidean space. As the rod has only a y component in K, its length there is $l^2 = y^2$. In K' the rod has both x and y components, and therefore its length is $l'^2 = x'^2 + y'^2$. Due to the invariance of the length $l^2 = l'^2$, we have

$$y^2 = x'^2 + y'^2 = y'^2(1+\tan^2\alpha) \ ,$$

where $\tan\alpha = x/y$. From here we determine y':

$$y' = \frac{y}{(1+\tan^2\alpha)^{1/2}}.$$

This relation shows that the height of the rod y' in K' is smaller than its height y in K. The reason is obvious – the K-observer measures the real, let us call it the *proper*, height of the rod, whereas the K'-observer determines its *apparent* height since the rod is inclined in K' and has not only a vertical (y') component, but also a horizontal

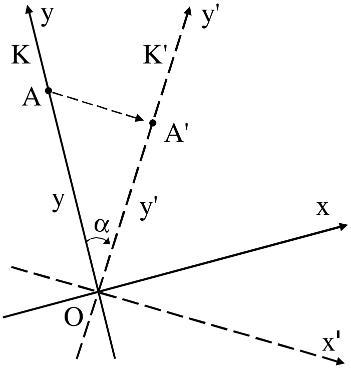


Fig. 4.17. Y 'dilation'

(x') component. This is an obvious fact in Euclidean space, but in spacetime, the *same* relation looks puzzling.

We have derived the relation (4.14) between the y components of the rod's length in K and K'. Now we can translate it into the relation between the time components of a four-dimensional rod (say, the worldtube of a clock) in two inertial reference frames S and S' in relative motion by using the substitution (4.4) and more specifically $y = ct$ and $\tan\alpha = \mathrm{i}\beta$:

$$ct' = \frac{ct}{(1-\beta^2)^{1/2}}\ ,$$

or in terms of time alone:

$$t' = \frac{t}{(1-\beta^2)^{1/2}}\ . \tag{4.15}$$

In order to understand the meaning of (4.15), assume that the inertial reference frames S and S' have at their origins two identical digital clocks. The time axes of S and S' are chosen along the clock worldlines as shown in Fig. 4.18. The two clocks are set to zero when they meet at event O.

An observer in S makes a simple measurement – he determines a short period of time starting at event O when the clock shows the zeroth second on its screen and ends at event A when the clock screen displays the fifth second. In other words the S-observer measures part of the spacetime length of a four-dimensional rod – the clock worldline – namely the length OA. As in Fig. 4.17, here too the rod has only

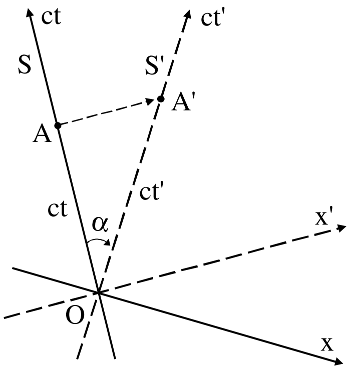


Fig. 4.18. Time dilation

'height' in S, i.e., only a time component. Let us call it *proper* time; hence the proper time measured by the S-observer is 5 s.

An observer in S' decides to determine the time component (the 'height') of that part of the four-dimensional rod of length OA in S' in order to compare the time components in S and S'. He projects the event A onto the event A' and finds that the time component (the 'height') of the four-dimensional rod is greater than t and is, say, $t' = 6$ s in S'. Unlike the Euclidean case where $y' < y$, in spacetime $t' > t$, as can be seen from (4.15) and Fig. 4.18. For this reason the described time effect is called time dilation. Here again it is clear why the S- and S'-observers disagree on what the time component of the worldline OA of the clock in S is. In S the clock worldline lies along the time axis and has only a time component ('height'); that is why the S-observer measures the proper length of the clock worldline, which we called proper time. In S' the worldline of the clock at rest in S is inclined and thus has both temporal and spatial components. That is why what the S'-observer measures is an apparent or dilated time.

If we compare Figs. 4.17 and 4.18, we see that they depict the *same* relation between the vertical lines in the two figures. So the time dilation effect turns out to be merely a manifestation of the fact that the worldline of a clock has only a time component ('height') in its rest frame, whereas in another inertial reference frame, which is in relative motion with respect to the clock's rest frame, the clock worldline is inclined and therefore has both temporal and spatial components there. In this way the radical research team obtain what they have been seeking – a prediction that can be experimentally tested. As we now know, experiment has confirmed the time dilation effect which, according to the radical research team, is experimental evidence to support the view that spacetime represents a real four-dimensional world. Judging by their experience with the traditional research team, the members of the radical team are well aware that there will be attempts to interpret the effects they predict in terms of the ordinary three-dimensional world. That is why they use an effective method for invading their opponents' territory. They would love to tell their opponents: "Fine. Let us assume you are right that the world is three-dimensional." Then they would show that none of the spacetime effects they predicted would be possible if the world were indeed three-dimensional. How this method works is demonstrated in Chap. 5.

Let us now return to the time dilation effect. As the Lorentz transformations (4.7) and (4.8) are the translated Euclidean rotation transformations (4.2) and (4.3), it follows that the time dilation effect can also be obtained from the Lorentz transformation (4.8). As the clock in question is at rest at the origin of S, the events O and A have the same x component $x = 0$ and (4.15) does follow from (4.8):

$$t' = \frac{t}{(1-\beta^2)^{1/2}} \ .$$

In order to determine the time component of the worldline OA in S′ the observer there waits until the event A falls in an instantaneous space corresponding to a given moment in S'. Then he notes that moment on the screen of the clock at rest at the origin of S'. One can visualize this by following the same procedure employed by the K'-observer in the Euclidean case. Imagine a line parallel to x' and start to move it upwards in S' until it reaches the event A. Then look at the event where the line intersects the time axis of S', i.e., where the line intersects the worldline of the clock at rest in S'. Put another way, the S'-observer *projects*[9] the A event onto event A' lying on the time axis of S'. He finds that the event A' corresponds to the 6th second of the time in S' and concludes that, while 5 s have elapsed between the events O and A for the S-observer, the apparent (dilated) duration between the same two events as determined in S' is 6 s.

There is nothing mystical about this result – simply the time component of the four-dimensional rod (the clock worldline) OA determined in S' is greater than the time component measured in S (exactly as the vertical y component of the three-dimensional rod in the Euclidean case is smaller in K'). In spacetime the effect is reversed due to the pseudo-Euclidean nature of spacetime. The time dilation effect has a natural explanation if spacetime is real, which means that the worldlines of the clocks at rest at the origins of S and S' are real four-dimensional objects. However, if spacetime were not real, the time dilation effect would not be just a puzzle – it would be impossible if the existence of the objects involved in this effect is regarded as absolute, as we will see in Chap. 5.

Before starting the discussion of another manifestation of the reality of spacetime, let us note that the time dilation effect can also be obtained through the invariance of the spacetime interval. The spacetime interval between events O and A determined in S is

[9] Two Lorentz transformations – active and passive – are distinguished in the books on relativity (see for instance [22, p. 49] and [41, p. 1140]). Throughout this book, only the passive Lorentz transformations will be used. In the passive view the *same* event of spacetime has different coordinates in different reference frames.

$$s^2 = c^2t^2 \ ,$$

since O and A have the same spatial coordinate $x = 0$. Here t is the proper time between O and A. In S', the spacetime interval between the same events is

$$s'^2 = c^2t'^2 - x'^2 \ .$$

Due to the invariance of the spacetime interval $s^2 = s'^2$, we have

$$c^2t^2 = c^2t'^2 - x'^2 = c^2t'^2(1 - \beta^2) \ ,$$

where we have taken into account the fact that $x'/t' = v$, v being the relative speed between S and S'. Finally,

$$t' = \frac{t}{(1 - \beta^2)^{1/2}} \ .$$

An important feature of the time dilation effect is that it is reciprocal. As shown in Fig. 4.19, the S-observer determines that the spacetime length OA of the worldline of the digital clock at rest at the origin of S is $s = ct$; that is, the *proper* time between events O and A is $t = 5$ s. In S', however, the spacetime length OA has both temporal and spatial components, which means that the S'-observer does not measure the proper time between events O and A along the S-clock. What he measures is the *apparent* or *dilated* time between the same events – he finds that event A' lying on the worldline of the S'-clock is *simultaneous*[10] with A and, since event A' is the S$'$-digital clock existing at the 6th second of its history, concludes that the duration between O and A as determined in S' is $t' = 6$ s.

The same procedure is employed by the S-observer when he wants to determine the duration between events O and B' in S; the proper time between these events as determined in S' is $t' = 5$ s. The S-observer finds that event B is simultaneous with B' and concludes that the duration between O and B' as measured in S is $t = 6$ s. Each of the observers measures five seconds in his own reference frame but when he looks at the measurement of the other observer determines that the time in the other reference frame appears dilated. Figure 4.19 demonstrates that nothing happens to the times of each of the observers. The scale along the worldlines of the two clocks is the *same* for both observers, which means that five seconds of the proper time

[10] The S$'$-observer determines that A' is simultaneous with A by noting that A falls in the instantaneous three-dimensional space corresponding to event A'. In other words, he *projects* event A onto event A'.

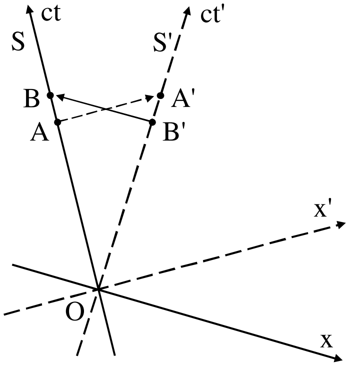


Fig. 4.19. Time dilation is reciprocal

in S is equal to five seconds of the proper time in S'. Put another way, proper time is *invariant*, since it is proportional to the spacetime length $s = ct$ of a time-like worldline.

The reciprocity of the time dilation effect provides an excellent opportunity to ask the fundamental question: Are the two clock worldlines real four-dimensional objects or are they nothing more than convenient graphical representations of the consequences of special relativity? We will address this question in more detail in Chap. 5, but one can try to see here whether the time dilation effect can be reciprocal if one assumes that the worldlines are not real four-dimensional objects, which means that each of the clocks exists only at *one* moment of its history.

4.5 Length Contraction in Euclidean Space and in Spacetime

We have derived the relation between the 'vertical' components of a given length as determined in two coordinate systems in Euclidean space and in two inertial reference frames in spacetime. It is now quite natural to ask what the relation is between the 'horizontal' components of a length in Euclidean space and in spacetime. As we did in the case of the time dilation effect, here too we will first derive that relation in Euclidean space and then translate it in order to obtain the corresponding relation in spacetime, known as length contraction. However, length contraction turns out to be more subtle than time dilation. Unfortunately, this subtlety has not been addressed in the books on special relativity.

In the time dilation effect two observers in relative motion measure the time component of the *same* time-like worldline, say the length OA in Fig. 4.18. One of the observers measures the *true* length of the worldline since the time axis of the observer's reference frame is chosen along the worldline, part of which is OA; therefore, for this observer OA has only a time component, which we called proper time. In the other observer's reference frame, however, OA has *both* temporal and spatial components. In the length contraction effect we would like to determine the length of the same rod as measured by two observers in relative motion. Assume that the rod is at rest in the inertial reference frame S and lies along the x axis. Its length $l = 1$ m there is called *proper* length. The rod has only a 'horizontal', i.e., spatial component in S. It does not have a temporal component since all of its parts exist *simultaneously* at any moment of the rod's history. In other words, according to the three-dimensionalist view, the entire length of the rod is given at once at only *one moment* of its history, i.e., the present moment; if the rod existed at more moments *at once* (not just at one) it would not be three-dimensional but four-dimensional, as shown in Fig. 4.1b for the case of a digital clock. Thus a three-dimensional object exists consecutively at all moments of its history but never exists at more than a single moment at once. An extended three-dimensional object is by definition *entirely given* at one moment of its history, which means that its parts exist *simultaneously* at that moment.

When an observer in another inertial reference frame S' in relative motion with respect to S (along the x axis) determines that the length of the same rod is shorter in S', he must also measure it at *one* moment in S'; that is, all parts of the rod must exist *simultaneously* in S' at any moment of its history. This means that the rod must not have a time component in S' either.[11] This requirement is stated in all derivations of the length contraction in books on relativity. What has not been addressed, however, is the fact that the *same* three-dimensional rod *cannot* exist in more than one reference frame. The rod is an *extended* three-dimensional object and therefore its three-dimensional parts exist *simultaneously* at a given moment. Hence, due to the relativity of simultaneity, the rod cannot exist in two reference frames in relative motion since *different* sets of events are simultaneous in the two frames.

[11] Here is the difference between time dilation and length contraction. As shown in Fig. 4.18, the time-like length OA has a spatial component in S'. In the length contraction effect the rod must have only spatial components in both S and S'.

The very fact that an observer in S' does measure the (contracted) length of the rod has only one explanation – the S′-observer does not measure the *same* three-dimensional rod; he measures a *different* three-dimensional object. This appears to be total nonsense since the rod is by definition *one*. We have seen that the radical research team cannot be impressed by apparent paradoxes. They have gotten accustomed to using the method of temporarily accepting the common view. So, they could say: "Let us assume what appears to be completely obvious – that two observers in relative motion measure the *same* rod." As all parts of the *extended* rod exist *simultaneously* at any moment of its history, it inevitably follows that relativity of simultaneity would be impossible, because both the S- and S'-observers would have the *same* set of simultaneous events – the same extended three-dimensional rod.

This apparent contradiction is encouraging for the radical research team. It clearly supports their view that spacetime is real. In a four-dimensional world the rod is a four-dimensional object – the rod's worldtube. The three-dimensional spaces of the S- and S'-observers intersect the worldtube at different places, as shown in Fig. 4.20, and the two three-dimensional cross-sections are interpreted by the S- and S'-observers as two different three-dimensional rods. This explains not only why the S- and S'-observers measure different three-dimensional rods, but also why their length is different – the three-dimensional spaces of S and S' intersect the rod's worldtube at different angles and the resulting cross-sections are therefore of different length.

It is true that the rod is by definition one object, but that object is the four-dimensional rod's worldtube. It is also true that the S- and S'-observers measure different three-dimensional rods, but these are merely different three-dimensional cross-sections of the rod's worldtube. The conclusion that, while measuring the same rod, two observers in relative motion measure different three-dimensional objects in fact follows directly from the basic assumption of the radical research team that the observers have *different* three-dimensional spaces and therefore different three-dimensional objects.

Now it is the traditional research team's turn not to be impressed by the conceptual analysis of the length contraction carried out by the radical research team. They tell their opponents that what a physicist should do is derive the length contraction from the Lorentz transformations and not bother about conceptual and interpretative questions. This is the moment the members of the radical team have been waiting for: "Excellent! It was precisely the application of the Lorentz trans-

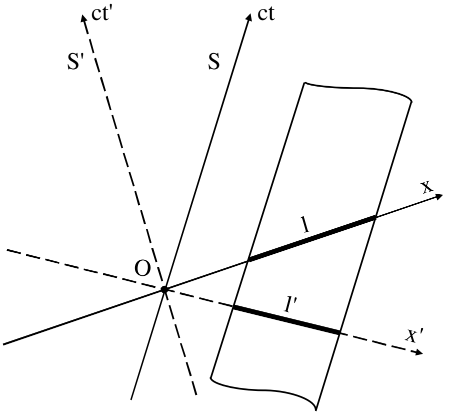


Fig. 4.20. The physical meaning of length contraction is rather counter-intuitive – two observers in relative motion do not measure the *same* three-dimensional rod, since the three-dimensional spaces of S and S' intersect the worldtube of the rod in two different three-dimensional cross-sections which each of the observers regards as his rod. That the observers measure different three-dimensional rods follows from the relativity of simultaneity – the rod is an *extended* object which means that its parts exist *simultaneously* at a given moment of the time of an observer; if two observers in relative motion measured the same rod, it would mean that simultaneity would be common for them and therefore absolute

formations that made us analyze the physical meaning of determining the length of three-dimensional objects in two reference frames in relative motion." Then they explain to their traditional colleagues what they mean. The rod is at rest in S and the coordinates x_P and x_Q of its end points are known in S (Fig. 4.21). This means that its length $l = x_Q - x_P$ is also known in S.

At the event O when the origins of S and S' coincide instantaneously, the S'-observer decides to determine the length of the rod in S'. As in the case of time dilation, he intends to carry out the transformation $S \rightarrow S'$, since he wants to use the *known* coordinates x_P and x_Q to obtain the unknown coordinates $x'_{p'}$ and $x'_{q'}$ of the end points of the rod in S'; he obtained the expression for time dilation in exactly the same way. The Lorentz transformation (4.7) for x and x' then gives

$$x'_{q'} - x'_{p'} = \frac{x_Q - x_P}{(1 - \beta^2)^{1/2}} \, . \tag{4.16}$$

Here we took into account the fact that events P and Q have the same time components in S. When the radical research team derived (4.16)

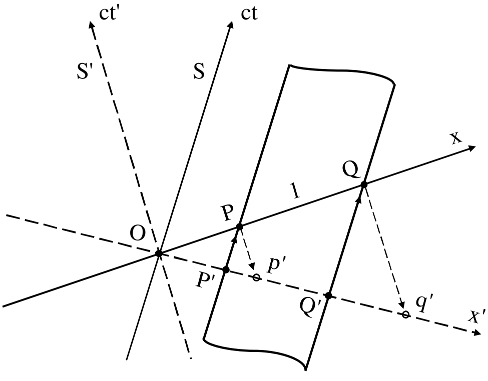


Fig. 4.21. Length contraction is more subtle than time dilation

and looked at the spacetime diagram shown in Fig. 4.21, they realized that their main hypothesis that spacetime is real was in trouble. As the rod is represented by its four-dimensional worldtube in spacetime, the three-dimensional spaces of S and S' intersect the worldtube at the three-dimensional cross-sections PQ and $P'Q'$, respectively. For this reason the researchers of the radical team expected the Lorentz transformation $S \rightarrow S'$ to project the events P and Q onto the events P' and Q'. However, $S \rightarrow S'$ projected the events P and Q onto p' and q'. If the distance $p'q'$ were regarded as the length l' of the rod in S' (which can be considered since p' and q' are *simultaneous* in S'), then measuring lengths in S and S' would be similar to the time dilation effect: the apparent length l' would be greater than the proper length l. Experiment would show later that, in reality $l' < l$, but in their analysis the researchers of the radical team could not use that evidence; they have been trying to derive predictions from their major hypothesis that spacetime is real, which could be tested experimentally.

The first thing the members of the radical research team check is whether they used the correct Lorentz transformation. The transformation $S \rightarrow S'$ does appear to be the one that should be used since it expresses the *unknown* coordinates of the end points of the rod in S' as a function of the *known* coordinates already determined in S. Moreover this transformation worked perfectly in the case of time dilation.[12] At this point they start to analyze the physical meaning of

[12] If the length PQ were part of a space-like worldline, then the Lorentz transformation $S \rightarrow S'$ would be the right transformation since the S-observer would measure the true length PQ, whereas the S'-observer would determine the 'horizontal' (spatial) component $p'q'$ of the length PQ. However, the length PQ of

measuring length, first in one reference frame and then in two frames in relative motion. As we have seen above this analysis led them to the conclusion that in *both* S and S' the rod has *only* spatial components. Although in the derivations of the length contraction effect in the books on relativity it is explicitly stated that the length of the rod is measured *simultaneously* in each of the frames S and S', it has not been explained that the physical meaning of this requirement is that the S- and S'-observers measure *different* three-dimensional rods. The researchers of the radical team challenge their traditional colleagues to find an error in their analysis. Readers are encouraged to try themselves, to understand why the radical researchers are so confident in their results.

Initially, the assumption that the Lorentz transformation $S \rightarrow S'$ is the correct one to use appeared to contradict the four-dimensionalist view and more specifically its corollary that the rod's worldtube is a real four-dimensional object. Now the fact that the correct[13] transformation $S \rightarrow S'$ does not adequately reflect the measurement of the length of a rod in two reference frames in relative motion constitutes a strong argument in favour of the reality of spacetime. As shown in Fig. 4.21, the S-observer regards the three-dimensional cross-section PQ as his rod, whereas the cross-section $P'Q'$ is the three-dimensional rod which is measured by the S'-observer. The Lorentz transformation that relates the two cross-sections is $S' \rightarrow S$, since it projects the events P$'$ and Q' onto P and Q, as shown in Fig. 4.21. The transformation $S' \rightarrow S$ for the x coordinates is

$$x = \frac{x' + \beta ct'}{(1-\beta^2)^{1/2}} \ . \tag{4.17}$$

Making use of this, we obtain

$$x_P - x_Q = \frac{x'_{Q'} - x'_{P'}}{(1-\beta^2)^{1/2}} \ .$$

From here we can determine the length of the rod $l' = x'_{Q'} - x'_{P'}$ in S':

$$x'_{Q'} - x'_{P'} = (x_Q - x_P)(1-\beta^2)^{1/2} \ ,$$

the rod (i.e., the three-dimensional rod itself) is not part of a space-like worldline (although the distance PQ is space-like); if this were the case the rod would not last in time but would appear at only one instance and then disappear forever.

[13] It is the correct transformation in the sense that it expresses the unknown coordinates of the rod in S' as a function of its known coordinates in S.

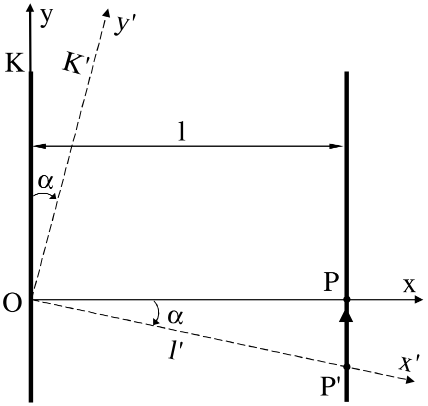


Fig. 4.22. Length ‘contraction’ in Euclidean space

or (taking into account the fact that $l = x_Q - x_P$)

$$l' = l(1 - \beta^2)^{1/2} . \tag{4.18}$$

We have obtained the contracted length (4.18) of the rod as determined in S' by employing the transformation $S' \to S$. The unusual[14] application of this transformation demonstrates that it links the lengths of two *different* three-dimensional rods, which only appears possible in a four-dimensional world (we return to this issue in Chap. 5).

Let us briefly follow the radical research team in their translation of the length contraction effect as defined in Euclidean space into the corresponding effect in spacetime. The strip defined by the two thick vertical lines in Fig. 4.22 has horizontal length l in the coordinate system K. The ‘horizontal’ length of the strip in K' is l'. The fact that the K- and K'-observers measure *different* lengths, which are different cross-sections of the x and x' axes and the strip, surprises no one in the Euclidean case. But if spacetime is real and the worldtubes of the physical bodies are real four-dimensional objects, then an analog of the Euclidean situation shown in Fig. 4.22 will inevitably be present when different observers in relative motion measure the length of the same physical body.

The Euclidean relation between l and l' is obvious:

$$l = l' \cos\alpha = \frac{l'}{(1 + \tan^2\alpha)^{1/2}} ,$$

[14] It is unusual since it transforms the *unknown* coordinates $x'_{P'}$ and $x'_{Q'}$ into the *known* coordinates x_P and x_Q.

which can be written as

$$l' = l(1 + \tan^2 \alpha)^{1/2}. \tag{4.19}$$

As seen from (4.19), in Euclidean space, the length contraction is in fact length dilation. This is not surprising. As discussed above, the pseudo-Euclidean nature of spacetime is responsible for such discrepancies.

We have obtained (4.19) directly, but the reader can check that it can be obtained, not by what appears to be the correct transformation $K \to K'$ [since it expresses the *unknown* coordinates of P' in K' in terms of the *known* coordinates of P in K; the coordinates of O are $(0, 0)$ in both K and K'], but by the transformation $K' \to K$. It is only the transformation $K' \to K$ that links the lengths OP and OP'.

The Euclidean analog of the length 'contraction' effect (4.19) can be translated into the corresponding spacetime effect again by making use of the substitution (4.4). Note that $l = \Delta x$ and $l' = \Delta x'$, which are replaced by $\mathrm{i}\Delta x$ and $\mathrm{i}\Delta x'$, and $\tan \alpha$ is replaced by $\mathrm{i}\beta$. The result is

$$\mathrm{i}\Big[\Delta x' = \Delta x(1 - \beta^2)^{1/2}\Big] \,.$$

Applying the rule to interpret only the real expressions in the brackets after the imaginary unit and writing this relation in terms of l and l', we obtain,

$$l' = l(1 - \beta^2)^{1/2} \,.$$

This result further supports the expectations of the members of the radical research team that, by translating simple geometrical relations in Euclidean space, they will be able to deduce predictions which can be used to test the reality of spacetime experimentally. And one day they would be delighted to read what Hermann Minkowski, another radical researcher, wrote [9, p. 76]:

> The whole universe is seen to resolve itself into ... world-lines, and I would fain anticipate myself by saying that in my opinion physical laws might find their most perfect expression as reciprocal relations between these world-lines.

4.6 The Twin Paradox in Euclidean Space and in Spacetime

The researchers of the radical team have succeeded in deducing two predictions – the time dilation and length contraction effects – which in their view are manifestations of the reality of spacetime. Both effects can be formulated only when two reference frames defining three-dimensional spaces and time directions are introduced. As spacetime itself is not *objectively* divided into space and time, these effects do not exist objectively – there are no separate three-dimensional spaces in spacetime that intersect the physical objects' worldtubes and thus give rise to time dilation and length contraction; there are only worldtubes there.[15]

In their search for relations between lines in Euclidean space that can be translated into the corresponding relations between worldlines in spacetime, the radical researchers have realized that the triangle inequality (the sum of the lengths of any two sides of a triangle is greater than the length of the third side) constitutes a relation between lines that does not need the introduction of coordinate systems or reference frames to exist. Figure 4.23a depicts a triangle in the ordinary Euclidean space. It is obvious that the sum of the lengths of the two sides DT and TM is greater than the length of the third side DM. In other words, in Euclidean space, the shortest distance between two points is along the straight line connecting the two points – in Fig. 4.23a the shortest distance between D and M is along the straight line DM.

Figure 4.23b shows a triangle in spacetime whose lines may be worldlines of physical objects. Assume that those worldlines represent two twins A and B. The straight worldline DM is the worldline of twin A, who does not change his state of motion, whereas the worldlines DT and TM belong to the worldline of twin B, who starts a round trip at the event of departure D, after some time turns back at event T, and meets his brother at event M.[16] It is clear from the spacetime diagram that the worldlines of the two twins between D and M are different. This means that different amounts of their proper times will have elapsed when they meet at M, since the lengths of the twins' worldlines are proportional to their proper times. The members of the radical research team have already developed a pseudo-Euclidean intuition

[15] However, if we *describe* spacetime in the ordinary three-dimensional language by introducing reference frames, the two effects can be experimentally tested since their very existence demonstrates that the objective world is not the three-dimensional world of our perceptions. If the objective world were indeed three-dimensional, there would be no such effects as time dilation and length contraction. It will be truly helpful for a genuine understanding of the arguments presented in this book if the reader tries to disprove every statement like this one.

[16] In Chap. 5, we will carry out a conceptual analysis of the twin paradox in an attempt to gain further insight into its physical meaning.

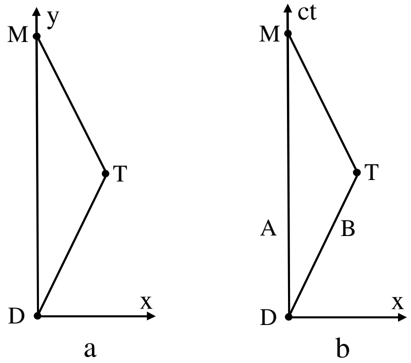


Fig. 4.23. The twin paradox in Euclidean space (**a**) and in spacetime (**b**)

and expect the triangle inequality in spacetime to be different from that in Euclidean space. Their intuition tells them that, in contrast to the Euclidean case, in spacetime, the longest path between two events might be along the straight worldline connecting them. As the twin paradox is of special importance to the radical researchers, since it is an *objective* and in this sense *absolute* effect (existing without the introduction of reference frames), they decide to carry out the calculations for the Euclidean triangle despite its obviousness and then:

- translate the expression obtained for the Euclidean space into the corresponding spacetime expression,
- follow the same procedure employed in the Euclidean case in order to analyze the effect in spacetime itself.

To compare the lengths of the sides of the triangle we will choose three coordinates systems,[17] K^1, K^2, and K^3, whose y axes are along the sides DM, DT, and TM, of the triangle, respectively, as shown in Fig. 4.24. In order to obtain the triangle inequality in Euclidean space, we will make use of the invariance of the Euclidean distance.

Let us determine the relation between the sides of the triangle in the coordinate system K^1. As the side DM has only a y component, its length l_{DM}^1 is obviously equal to y_{DM}^1. The length of the side DT

[17] Note that we need the coordinate systems only to calculate the relation between the sides of the triangle. The fact that the sum of the lengths of the two sides of the triangle is greater than the length of the third side is independent of the introduction of any coordinate systems. Recall that the y 'dilation' and length 'contraction' effects in Euclidean space can be defined only in terms of two coordinate systems; they have no independent meaning in Euclidean space without the introduction of coordinate systems.

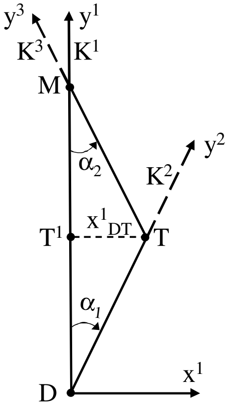


Fig. 4.24. The twin paradox in Euclidean space

has both x and y components in K^1:

$$l^1_{DT} = \left[(y^1_{DT^1})^2 + (x^1_{DT})^2\right]^{1/2} = y^1_{DT^1}(1 + \tan^2\alpha_1)^{1/2}\ .$$

Taking into account the fact that $l^1_{DT^1} = y^1_{DT^1}$, we can write

$$l^1_{DT^1} = \frac{l^1_{DT}}{(1 + \tan^2\alpha_1)^{1/2}}\ .$$

In the same way, we obtain

$$l^1_{T^1M} = \frac{l^1_{TM}}{(1 + \tan^2\alpha_2)^{1/2}}\ .$$

As $l^1_{DM} = l^1_{DT^1} + l^1_{T^1M}$, the relation between the lengths of the three sides of the triangle is

$$l^1_{DM} = \frac{l^1_{DT}}{(1 + \tan^2\alpha_1)^{1/2}} + \frac{l^1_{TM}}{(1 + \tan^2\alpha_2)^{1/2}}\ . \tag{4.20}$$

The triangle inequality

$$l^1_{DM} \leq l^1_{DT} + l^1_{TM}$$

is contained in (4.20), since the first term in (4.20) is smaller than l^1_{DT} (as $\tan^2\alpha_1 > 0$ for $\alpha_1 \neq 0$) and the second term is smaller than l^1_{TM} (as $\tan^2\alpha_2 > 0$ for $\alpha_2 \neq 0$). The triangle inequality becomes even more evident if we choose $\alpha_1 = \alpha_2 = \alpha$:

$$l^1_{DM} = \frac{l^1_{DT} + l^1_{TM}}{(1 + \tan^2 \alpha)^{1/2}} \ , \tag{4.21}$$

which can be written as

$$l^1_{DM}(1 + \tan^2 \alpha)^{1/2} = l^1_{DT} + l^1_{TM} \ .$$

If $\alpha = 0$, the inequality becomes an equality, since in this case $l^1_{DT} = l^1_{DT^1}$ and $l^1_{TM} = l^1_{T^1 M}$.

For the purpose of translating the relation (4.21) into the spacetime relation, it can be written in terms of the y components of the sides of the triangle in the three coordinate systems. In order to do this, we use the invariance of the Euclidean length. In K^1 the lengths of the sides DT and TM are l^1_{DT} and l^1_{TM}, respectively. In K^2 the length of DT is $l^2_{DT} = y^2_{DT}$ and since $l^1_{DT} = l^2_{DT}$ we can write $l^1_{DT} = y^2_{DT}$. In K^3 the length of the side TM is $l^3_{TM} = y^3_{TM}$ and therefore $l^1_{TM} = y^3_{TM}$. Now we can rewrite (4.21) using only the y components of the sides of the triangle in K^1, K^2, and K^3:

$$y^1_{DM} = \frac{y^2_{DT} + y^3_{TM}}{(1 + \tan^2 \alpha)^{1/2}} \ . \tag{4.22}$$

Note that (4.22) relates the *proper* heights of the three sides as determined in their coordinate systems where they have only vertical or y components; these are not the apparent heights or y *projections* as in the y 'dilation' in Euclidean space.

The worldlines of twins A and B are depicted in Fig. 4.25. As in the Euclidean case, here too we introduce three reference frames – S^1 is associated with twin A, S^2 with twin B on his way toward the turning point (event T) (i.e., with the part of his worldline DT), and S^3 with twin B on his way back to twin A (i.e., with the worldline TM). In S^1 the time axis is along twin A's worldline DM, which means that its spacetime length there is $s^1_{DM} = ct^1_{DM}$, where t^1_{DM} is the *proper* time between the events D and M. In S^2 the spacetime length of twin B's worldline DT is $s^2 = ct^2_{DT}$, where t^2_{DT} is the *proper* time between the events D and T. In S^3 the spacetime length of the other part of twin B's worldline TM is $s^3 = ct^3_{TM}$, where t^3_{TM} is the *proper* time between the events T and M.

To translate the Euclidean relation (4.22) between the sides of the triangle in Fig. 4.24, we use the substitutions $y \rightarrow ct$ and $\tan \alpha \rightarrow \mathrm{i}\beta$ again:

$$ct^1_{DM} = \frac{ct^2_{DT} + ct^3_{TM}}{(1 - \beta^2)^{1/2}} \ ,$$

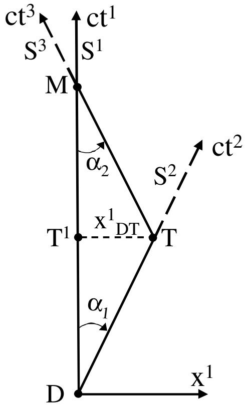


Fig. 4.25. The twin paradox in spacetime

or in terms of the proper times t^1_{DM}, t^2_{DT}, and t^3_{TM} alone,

$$t^1_{DM} = \frac{t^2_{DT} + t^3_{TM}}{(1-\beta^2)^{1/2}} \ . \tag{4.23}$$

Note that this relation is between *proper* times, not between proper and dilated times, as was the case in the time dilation effect. The time effect (4.23), known as the twin paradox, shows that when the twins meet at event M their clocks will show different times.[18] Twin B's clock will show that the duration of his round trip was $t^2_{DT} + t^3_{TM}$, whereas twin B's journey lasted *longer* according to twin A's clock and its duration is given by (4.23). This is an *absolute* (not a reciprocal) effect since both twins agree when they meet at M that twin B is younger.

As the researchers of the radical team expected, the triangle inequality in spacetime is reversed:

$$t^1_{DM} \geq t^2_{DT} + t^3_{TM} \ .$$

To see this better let us write (4.23) in the form

$$t^1_{DM}(1-\beta^2)^{1/2} = t^2_{DT} + t^3_{TM} \ . \tag{4.24}$$

It is obvious that t^1_{DM} is greater than the sum of t^2_{DT} and t^3_{TM}; only when the twins are at rest with respect to each other (i.e., when

[18] This is the standard version of the twin paradox when the relative speed between A and B during both parts of B's journey (when he recedes from A and returns back to A) is the same.

$\beta = v/c = 0$) is t^1_{DM} equal to the sum of t^2_{DT} and t^3_{TM}. The triangle inequality (4.23) shows that, in the pseudo-Euclidean geometry of spacetime, the longest spacetime distance (i.e., the longest proper time in the case of time-like worldlines) between two events is along the straight worldline connecting the events. This is a completely counterintuitive result but the radical team would be rewarded for their intellectual bravery when experiment confirmed the twin paradox effect.

Let us now apply the same procedure as in the Euclidean case and derive the relation between the proper times of the twins in spacetime itself. In the inertial reference frame S^1, where twin A is at rest, the spacetime length of twin A's worldline between the events D and M is $s^1_{DM} = ct^1_{DM}$. The spacetime length of twin B's worldline between events D and T has both temporal and spatial components in S^1 and is therefore

$$s^1_{DT} = \left[(ct^1_{DT^1})^2 - (x^1_{DT})^2\right]^{1/2} = ct^1_{DT^1}(1 - \beta_1^2)^{1/2} \; , \qquad (4.25)$$

where $\beta_1 = v_1/c$ and $v_1 = x^1_{DT}/t^1_{DT^1}$ is the relative speed between the twins during the first part of twin B's journey, and v_1 corresponds to the angle α_1 between the twins' worldlines at event D. Special attention should be paid to the time $t^1_{DT^1}$. On the one hand, it is the proper time between events D and T^1, where T^1 is simultaneous with T in S^1. On the other hand, however, it is the time between D and T as measured in S^1 (it is measured between D and T^1, since T^1 is simultaneous with T); that is, it is the dilated time corresponding to the proper time t^2_{DT} between the events D and T measured in S^2.

Due to the invariance of the spacetime distance $s^1_{DT} = s^2_{DT}$ and since $s^2_{DT} = ct^2_{DT}$, where t^2_{DT} is the proper time between events D and T, it follows that $s^1_{DT} = ct^2_{DT}$. Then we can take this into account in (4.25):

$$t^2_{DT} = t^1_{DT^1}(1 - \beta_1^2)^{1/2} \; ,$$

which can be written as[19]

$$t^1_{DT^1} = \frac{t^2_{DT}}{(1 - \beta_1^2)^{1/2}} \; . \qquad (4.26)$$

The spacetime length of twin B's worldline between events T and M also has both temporal and spatial components in S^1:

[19] You can recognize here the time dilation effect mentioned above – t^2_{DT} is the proper time between the events D and T, whereas $t^1_{DT^1}$ is the dilated time between the same events.

$$s^1_{TM} = \left[(ct^1_{T^1M})^2 - (x^1_{TM})^2\right]^{1/2} = ct^1_{T^1M}(1-\beta_2^2)^{1/2} \,, \qquad (4.27)$$

where $\beta_2 = v_2/c$ and $v_2 = x^1_{MT}/t^1_{MT^1}$ is the relative speed between the twins during the second part of twin B's journey ($x^1_{DT} = x^1_{TM}$ since the events D and M have the same x-coordinate in S^1), and v_2 corresponds to the angle α_2 between the twins' worldlines at event M.

Using once again the invariance of the spacetime distance $s^1_{TM} = s^3_{TM}$ and since $s^3_{TM} = ct^3_{TM}$, where t^3_{TM} is the proper time between T and M, we have $s^1_{TM} = ct^3_{TM}$. Then (4.27) becomes

$$ct^3_{TM} = ct^1_{T^1M}(1-\beta_2^2)^{1/2} \,,$$

which can be rearranged as

$$t^1_{T^1M} = \frac{t^3_{TM}}{(1-\beta_2^2)^{1/2}} \,. \qquad (4.28)$$

As the proper time t^1_{DM} between D and M is equal to the sum of the proper times $t^1_{DT^1}$ and $t^1_{T^1M}$ between DT^1 and T^1M, respectively, we have

$$t^1_{DM} = \frac{t^2_{DT}}{(1-\beta_1^2)^{1/2}} + \frac{t^3_{TM}}{(1-\beta_2^2)^{1/2}} \,.$$

We have obtained the relation between the proper times of the twins between the events D and M. For $\beta_1 = \beta_2 = \beta$, this relation becomes

$$t^1_{DM} = \frac{t^2_{DT} + t^3_{TM}}{(1-\beta^2)^{1/2}} \,, \qquad (4.29)$$

which coincides with (4.23) obtained by translating the corresponding Euclidean relation.

Here too it will be helpful for a genuine understanding of the meaning of the twin paradox to assume that the twins' worldlines are *not* real four-dimensional objects and to see whether this effect would be possible; we do this in Chap. 5.

4.7 Addition of Velocities

Let us now see whether the Lorentz transformations (4.11) are consistent with the consequence of the relativity principle that the speed of light is constant. So we have to verify whether a particle moving at the speed of light c in the frame S will move at the same speed in S' as well.

Let the speed of a particle moving along the x axis in S be $w^x = \mathrm{d}x/\mathrm{d}t$; in S' the particle's speed is $w^{x'}$. Using the Lorentz transformations (4.11) for $\mathrm{d}x'$ and $\mathrm{d}t'$, we have

$$w^{x'} = \frac{\mathrm{d}x'}{\mathrm{d}t'} = \frac{(\mathrm{d}x - \beta c\mathrm{d}t)\,/\,(1-\beta^2)^{1/2}}{(\mathrm{d}t - \beta\mathrm{d}x/c)\,/\,(1-\beta^2)^{1/2}} \ .$$

Dividing both numerator and denominator by $\mathrm{d}t$ and taking into account that $\beta = v/c$, we find

$$w^{x'} = \frac{w^x - v}{1 - vw^x/c^2} \ . \tag{4.30}$$

This is the relativistic expression for addition of velocities which replaces the classical rule $w^{x'} = w^x \pm v$.

Now assume that a particle is moving with speed $w^x = c$ in S. In S' its speed will be

$$w^{x'} = \frac{c - v}{1 - v/c} = c \ .$$

Therefore a particle moving at the velocity of light c in one inertial reference frame will move at c with respect to *all* inertial frames.

4.8 The Metric of Spacetime

Consider two events in spacetime. They can be connected by a displacement four-vector $\Delta\boldsymbol{x}$ whose components are:

$$\Delta x^\alpha = \{\Delta x^0, \Delta x^1, \Delta x^2, \Delta x^3\} \ . \tag{4.31}$$

As the magnitude of the displacement four-vector is the spacetime distance between the two events, it follows that the scalar product of the displacement four-vector with itself should be equal to the square of the spacetime distance Δs between the two events:

$$\Delta s^2 = \Delta\boldsymbol{x}\boldsymbol{\cdot}\Delta\boldsymbol{x} \ . \tag{4.32}$$

If we compare (4.32) with the spacetime interval

$$\Delta s^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2 \ , \tag{4.33}$$

we see that the right-hand side of (4.33) does resemble the expression for a scalar product, except that not all signs are the same. We know that this discrepancy is caused by the pseudo-Euclidean geometry of spacetime. So let us make use of this similarity and write the spacetime interval in a more compact form by employing the usual summation convention that there is always a summation over repeated upper and lower indices:

$$\Delta s^2 = \eta_{\alpha\beta}\Delta x^\alpha \Delta x^\beta \ , \tag{4.34}$$

where

$$\eta_{\alpha\beta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{4.35}$$

is called the *metric tensor* or simply the *metric* of the pseudo-Euclidean Minkowski spacetime. (As usual, all Greek letters run from 0 to 3.) It is evident that by taking into account (4.35), the compact expression for the spacetime interval (4.34) coincides with its explicit form (4.33). The metric (4.35) will allow us to calculate scalar products of four-vectors.

4.9 On Proper and Coordinate Time

Ordinary particles, whose velocities are smaller than c, follow time-like worldlines in spacetime. As we have seen above the length of a time-like worldline is proportional to the *proper* time τ of the corresponding particle. That is why the worldline of a clock can be regarded as a time ruler in spacetime. As we can use the proper time to parametrize a time-like worldline, we can express all four coordinates x^α of a particle as a function of τ:

$$x^\alpha = x^\alpha(\tau)$$

or

$$\begin{aligned} x^\alpha &= \left\{x^0(\tau), x^1(\tau), x^2(\tau), x^3(\tau)\right\} \\ &= \left\{ct(\tau), x(\tau), y(\tau), z(\tau)\right\} \ . \end{aligned} \tag{4.36}$$

Consider two events A and B lying on the worldline of a particle, as shown in Fig. 4.26. In the inertial reference frame S in which the particle is at rest, the spacetime length of the worldline between these events is the proper time multiplied by c:

$$\Delta s^2_{AB} = c^2 \Delta\tau^2_{AB} \ .$$

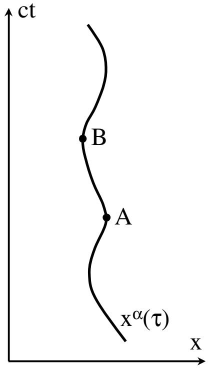


Fig. 4.26. Proper length of a time-like worldline

As the proper time τ is proportional to the spacetime distance, which is an invariant, it is clear that the proper time itself is an *invariant*. This is not surprising since the length of a worldline (like the length in the ordinary three-dimensional space) should not depend on the way it is measured.

In another frame S' in relative motion with respect to S, the same spacetime length is

$$\Delta s'^2_{AB} = c^2 \Delta t'^2_{AB} - \Delta x'^2_{AB} - \Delta y'^2_{AB} - \Delta z'^2_{AB} \ .$$

As the spacetime length is invariant, we have $\Delta s^2_{AB} = \Delta s'^2_{AB}$ and therefore,

$$\Delta \tau_{AB} = \Delta t_{AB} \left(1 - \beta^2\right)^{1/2} \ , \tag{4.37}$$

where

$$\beta^2 = \frac{\Delta x'^2_{AB} + \Delta y'^2_{AB} + \Delta z'^2_{AB}}{c^2 \Delta t^2_{AB}} = \frac{v^2}{c^2} \ .$$

We have derived (4.37) again in order to emphasize the similarities and differences between the proper time τ and the coordinate time t. In (4.37), the proper time $\Delta\tau_{AB}$ is proportional to the proper length $\Delta s_{AB} = c\Delta\tau_{AB}$ of the worldline between the events A and B. So $\Delta\tau_{AB}$ is a time period (between the events A and B as measured in S), but an invariant one. On the other hand, Δt_{AB} is also a time period between the same two events but measured in the frame S' with respect to which the particle moves. As Δt_{AB} is a *projection* of the time-like length AB onto the time axis of S', it is not an invariant period of time. The reason is that, in different reference frames whose time axes are not parallel, the projection will result in *different* coordinate time

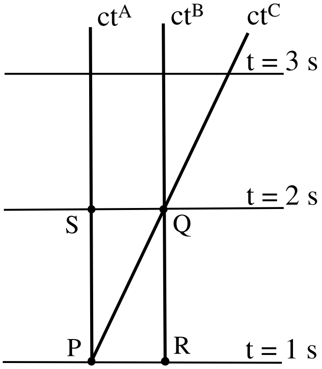


Fig. 4.27. The proper times of observers A, B, and C are measured along their worldlines. The horizontal lines of simultaneity define coordinate time in the inertial frame of A and B

periods Δt of the same spacetime distance AB, i.e., of the same proper time τ. If we put $v = 0$ in (4.37), which means that (4.37) is written in S, we see that, in an inertial reference frame, proper and coordinate times coincide: $\Delta\tau_{AB} = \Delta t_{AB}$.

The difference between proper and coordinate times can be further illustrated by considering three observers A, B, and C whose worldlines are shown in Fig. 4.27. Observers A and B are at rest with respect to each other, whereas C is moving uniformly relative to them. Each of the observers has a clock and measures his proper time along his worldline. So, proper time is measured with a *single* clock since the length of a time-like worldline is proportional to the proper time. As A and B are at rest relative to each other, they share the same inertial reference frame. Therefore they have common simultaneity. The lines of simultaneity corresponding to three moments of time are depicted in Fig. 4.27. These lines define *coordinate* time in the inertial reference frame of A and B – we can think of the lines of simultaneity as corresponding to each second of the coordinate (global) time in the inertial frame.

Imagine that observer C is performing an experiment which starts at event P and ends at event Q. So the proper time between P and Q is τ^C_{PQ} and is measured with only one clock. As shown in Fig. 4.27, the relativistically dilated time t_{PQ} is measured with *two* clocks – the beginning of the experiment P is measured with A's clock, whereas the end of the experiment is measured with B's clock. As any measurement with distant clocks involves coordinate time, the relativistically dilated time t_{PQ} is also expressed in terms of the coordinate time of the inertial reference frame in which A and B are at rest. According to A and B, the experiment performed by C lasts 1 s of the coordinate time of A and B; its proper time measured by C is shorter.

To see the different roles of coordinate and proper times, note that $t_{PQ} = t_{PS} = t_{RQ}$. On the one hand, events P and R are simultaneous – they lie on the line of simultaneity $t = 1$ s. On the other hand, events Q and S are simultaneous – they lie on the line of simultaneity $t = 2$ s. Therefore, the coordinate time between any event lying on the line $t = 1$ s and any event on the line $t = 2$ s is 1 s. It is evident from Fig. 4.27 that coordinate and proper time coincide in inertial reference frames – the proper times τ^A_{PS} and τ^B_{RQ} measured along the worldlines of A and B are equal to the coordinate times t_{PS} and t_{RQ}, respectively.

The lines of simultaneity in an inertial reference frame correspond to the instantaneous three-dimensional spaces belonging to different moments of time in the inertial frame. These spaces are parallel, as shown in Fig. 4.27. The spaces that correspond to different moments of the time of a non-inertial reference frame, however, are not parallel, as shown in Fig. 4.3. This means that coordinate and proper times do not coincide in non-inertial reference frames. In Fig. 4.28, 1 s coordinate time corresponds to different proper times for two observers A and B who are at rest in the non-inertial frame – for A the proper time is τ^A_{MN}, whereas for B it is τ^B_{OP}.

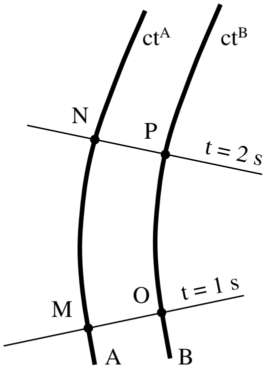


Fig. 4.28. The proper times of observers A and B are measured along their worldlines. The nearly horizontal lines of simultaneity define coordinate time in the non-inertial frame of A and B

4.10 Four-Velocity, Four-Momentum, and Relativistic Mass

When the motion of a particle is represented in the ordinary three-dimensional space by its trajectory, the tangent vector at a given point is the velocity of the particle at that point. We have a similar situation in spacetime. As in Fig. 4.29, the four-vectors $\boldsymbol{u}$ which are tangent to the time-like worldline of a particle can be regarded as its four-velocity, corresponding to different events of the worldline.

The components u^α of the four-velocity $\boldsymbol{u}$ are the derivatives of the four coordinates x^α of the particle with respect to its proper time:

$$u^\alpha = \frac{\mathrm{d}x^\alpha}{\mathrm{d}\tau} \,. \tag{4.38}$$

The explicit expressions for the components u^α of the four-velocity are

$$u^0 \equiv u^t = \frac{\mathrm{d}x^0}{\mathrm{d}\tau} = \frac{c\mathrm{d}t}{\mathrm{d}\tau} = \frac{c}{(1-\beta^2)^{1/2}} \,, \tag{4.39}$$

where

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \frac{1}{(1-\beta^2)^{1/2}}$$

is the relation between coordinate and proper time as seen from (4.37). The x component of the four-velocity is

$$u^1 \equiv u^x = \frac{\mathrm{d}x}{\mathrm{d}\tau} = \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau} = \frac{v^x}{(1-\beta^2)^{1/2}} \,, \tag{4.40}$$

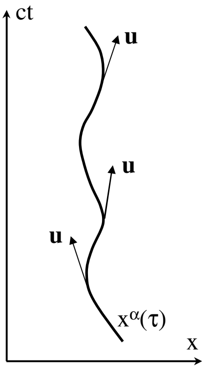


Fig. 4.29. The four-velocity of a particle is tangent to its worldline

where $\mathrm{d}x/\mathrm{d}t = v^x$ is the x component of the ordinary (relative) velocity of the particle with respect to the inertial reference frame in which it is determined. The components u^y and u^z have the same form as u^x.

As seen from (4.39), the time component of the four-velocity cannot be zero in any frame of reference, whereas the spatial components (4.40) of $\boldsymbol{u}$ become zero when the ordinary velocity of the particle is zero. As the four-velocity vector $\boldsymbol{u}$ is tangent to the worldline of a particle, $\boldsymbol{u}$ has only a time component in the reference frame where the particle is at rest, i.e., in its rest frame, since the four-velocity vector lies along the time axis of the frame.

Using the notation

$$\gamma = (1-\beta^2)^{-1/2} = \left(1 - \frac{v^2}{c^2}\right)^{-1/2} = \left[1 - \frac{(v^x)^2 + (v^y)^2 + (v^z)^2}{c^2}\right]^{-1/2} ,$$

we can write the components of the four-velocity as

$$u^\alpha = \left\{u^0, u^1, u^2, u^3\right\} = \left\{\gamma c, \gamma v^x, \gamma v^y, \gamma v^z\right\} , \tag{4.41}$$

or

$$u^\alpha = \left\{\gamma c, \gamma v\right\} ,$$

where $v = \{v^x, v^y, v^z\}$. By using the scalar product of the four-velocity vector with itself, we obtain the important result

$$\boldsymbol{u} \cdot \boldsymbol{u} = \eta_{\alpha\beta} u^\alpha u^\beta = (u^0)^2 - (u^1)^2 - (u^2)^2 - (u^3)^2 = c^2 . \tag{4.42}$$

If the four-velocity is normalized by choosing $c = 1$ in (4.42), then $\boldsymbol{u}$ turns out to be a unit time-like four-vector.

The four-velocity vector allows us to define the four-momentum of a particle whose so-called *rest* mass m is measured in its rest frame:

$$\boldsymbol{p} = m\boldsymbol{u} . \tag{4.43}$$

By taking into account (4.42), we can write for the scalar product of the four-momentum with itself:

$$\boldsymbol{p}^2 = \boldsymbol{p} \cdot \boldsymbol{p} = m^2 \boldsymbol{u} \cdot \boldsymbol{u} = m^2 c^2 . \tag{4.44}$$

We can easily write the components p^α of the four-momentum, since we know the components (4.41) of the four-velocity:

$$p^0 = \frac{mc}{(1-\beta^2)^{1/2}} , \tag{4.45}$$

and

$$p = \frac{mv}{(1-\beta^2)^{1/2}} \ . \tag{4.46}$$

Therefore

$$p^\alpha = \Big\{\gamma mc, \gamma mv\Big\} \ . \tag{4.47}$$

The spatial component of the four-momentum is the usual (classical) momentum mv with the relativistic correction γ. As can be seen from (4.46), for small velocities $v/c \ll 1$, it does reduce to the classical momentum:

$$p = mv + \cdots \ .$$

The physical meaning of the time component of the four-momentum is not immediately obvious. Let as see whether we can gain some insight from the expression (4.45) for p^0 for small velocities:

$$\begin{aligned} p^0 &= mc\left(1 + \frac{1}{2}m\frac{v^2}{c^2} + \cdots\right) = mc + \frac{1}{2}m\frac{v^2}{c} + \cdots \\ &= \frac{1}{c}\left(mc^2 + \frac{1}{2}mv^2 + \cdots\right) \ . \end{aligned} \tag{4.48}$$

The second term in brackets in (4.48) is the usual kinetic energy, but the first term is a new relativistic energy. Let us call it the *rest* energy of a particle, since it alone remains in the particle's rest frame, i.e., when the velocity in (4.48) is zero. If we denote the total energy of the particle for small velocities by

$$E = mc^2 + \frac{1}{2}mv^2 + \cdots \ ,$$

the expression for p^0 becomes

$$p^0 = \frac{E}{c} \ .$$

Now since p^0 cannot have different expressions for small and high velocities, its general expression is the same, but the energy in this case is obviously

$$E = \frac{mc^2}{(1-\beta^2)^{1/2}} \ . \tag{4.49}$$

This is the total relativistic energy of a particle moving with any velocity $v < c$. In its rest frame, a particle's energy reduces to its rest energy:

$$E = mc^2 \ .$$

In terms of the total relativistic energy the components of the four-momentum are

$$p^\alpha = \Big\{E/c, p\Big\} = \Big\{\gamma mc, \gamma mv\Big\} \ . \tag{4.50}$$

The physical meaning of the time component of the four-momentum now becomes clear – it is proportional to the energy of the particle. Therefore, two seemingly different dynamical characteristics of a particle, its energy and momentum, turn out to be different components of the same four-vector in spacetime. That is why the four-momentum is also called the *energy–momentum* vector.

Using (4.44) and (4.50), we can find the relativistic relation between E and p by calculating the scalar product of the four-momentum with itself:

$$\boldsymbol{p} \cdot \boldsymbol{p} = \eta_{\alpha\beta} p^\alpha p^\beta = (p^0)^2 - p^2 = m^2 c^2 \ .$$

As $p^0 = E/c$, we have

$$E^2 = m^2 c^4 + p^2 c^2 \ . \tag{4.51}$$

Let us now address the question of the so called relativistic increase of the mass, which has recently become controversial. We can write (4.46) and (4.49) in the form

$$p = \frac{mv}{(1-\beta^2)^{1/2}} = m(v)v \ ,$$

$$E = \frac{mc^2}{(1-\beta^2)^{1/2}} = m(v)c^2 \ ,$$

where

$$m(v) = \frac{m}{(1-\beta^2)^{1/2}} = \frac{m}{(1-v^2/c^2)^{1/2}} \tag{4.52}$$

is the relativistic mass of a particle, which is function of the particle velocity.

There have been recent objections against using the concept of relativistic mass. The argument goes as follows [21, pp. 250–251]:

> The concept of 'relativistic mass' is subject to misunderstanding [...]. First, it applies the name mass – belonging to the magnitude of a 4-vector – to a very different concept, the time component of a 4-vector. Second, it makes increase of energy of an object with velocity or momentum appear to be connected with some change in internal structure of the object. In reality, the increase of energy with velocity originates not in the object but in the geometric properties of spacetime itself.

It is true that the magnitude of the four-momentum is proportional to the rest mass of a particle as seen from (4.44):

$$|\boldsymbol{p}| = mc \, .$$

The time component of the four-momentum

$$p^0 = \frac{m}{(1 - v^2/c^2)^{1/2}} = m(v)c$$

is proportional to the relativistic mass $m(v)$. So it does appear that the rest (proper) mass m is proportional to the magnitude of a four-vector, whereas the relativistic mass $m(v)$ is a component of a four-vector. However, the situation is precisely the same with respect to proper and coordinate times. As seen from (4.32), the square of the spacetime distance Δs^2 between two events lying on a time-like worldline is equal to the scalar product $\Delta\boldsymbol{x}\cdot\Delta\boldsymbol{x}$ of the displacement four-vector $\Delta\boldsymbol{x}$ connecting the two events. In other words, the magnitude of the displacement vector is equal to the spacetime distance:

$$|\Delta\boldsymbol{x}| = \Delta s \, .$$

As $\Delta s = c\Delta\tau$, the magnitude of $\Delta\boldsymbol{x}$ is proportional to the proper time $\Delta\tau$ between the events connected by the displacement vector:

$$|\Delta\boldsymbol{x}| = c\Delta\tau \, .$$

Therefore, the magnitude of the four-vector $\Delta\boldsymbol{x}$ is proportional to the proper time $\Delta\tau$. On the other hand, however, coordinate time is the zeroth (time) component $\Delta x^0 = c\Delta t$ of the displacement four-vector $\Delta\boldsymbol{x}$ in (4.31).

So, if we cannot talk about relativistic mass, by the same argument we should talk only about proper time, which is an invariant, and deny the name ‘time’ to the coordinate time. Proper time is an invariant, but time can also change relativistically. The same holds for the mass as well. This becomes even more evident from the very definition of mass as the measure of the *resistance* a particle offers to its acceleration or, in the framework of relativity, as the measure of the resistance a particle offers when deviated from its geodesic path. That resistance is different in different reference frames with respect to which the particle moves with different velocities. Therefore the particle mass should also differ in different frames. It should be stressed that the resistance arises *in* the particle; it does not come from the geometric properties of spacetime. It is spacetime that determines the shape of a geodesic worldline, but it is the particle that resists when prevented from following a geodesic path. We have a proof that the resistance does not originate in the geometry of spacetime – a particle whose worldtube is deviated from its geodesic shape offers the *same* resistance in flat and curved spacetime as the equivalence of inertial and passive gravitational masses shows.

4.11 Summary

The main idea of this chapter has been to exploit the similarities between Euclidean space and spacetime. As the members of the radical research team wanted to find a way to test their major hypothesis – that spacetime is real – they found an extremely simple but effective way to deduce predictions from the spacetime hypothesis. If spacetime is a real four-dimensional space then relations between lines in Euclidean space should have analogs in spacetime. And indeed, when the radical research team 'translated' the Euclidean relations into the corresponding spacetime relations, it turned out that the consequences of special relativity are, in fact, manifestations of the four-dimensionality of spacetime.

At this point some might object that this proves nothing – what the radical research team did was an interesting exercise, but just an exercise, which does not force us to change our views on the world. What we have seen in this chapter is that, if the world is four-dimensional, there will be manifestations of its four-dimensionality which coincide with the consequences of special relativity. In Chap. 5, I will argue that those consequences would be impossible if the world were three-dimensional.

5 Relativity and the Dimensionality of the World: Spacetime Is Real

In the two previous chapters we have been following the radical research team in their analysis of Galileo's principle of relativity. Its members employed an improved version of Galileo's method of scientific inquiry and managed not to be bothered by the counter-intuitive nature of their conclusions. Now we know that the predictions derived by the radical research team to test the reality of spacetime coincide with the predictions of special relativity. Experiment has confirmed all consequences of special relativity, but few physicists and philosophers have regarded that confirmation as a proof of the existence of spacetime. There are two main reasons for this situation:

- the fact that in 1905 Einstein formulated the special theory of relativity in terms of our ordinary three-dimensional language,
- Minkowski's 1908 four-dimensional formulation of special relativity, in which he united space and time into what we now call Minkowski spacetime, has been regarded by most physicists and philosophers as merely a mathematical tool without an objective counterpart.

In this chapter we will examine the consequences of special relativity to see whether they themselves allow different interpretations of the nature of spacetime.[1] We will analyze four kinematic relativistic effects – relativity of simultaneity, time dilation, length contraction, and the twin paradox – by asking a simple question: Are these effects possible if Minkowski spacetime is not real? Obviously this question is equivalent to: Are those effects possible if the world is three-dimensional? I will argue that, not only would the kinematic consequences of special relativity be impossible if the world were three-dimensional, but

[1] As we will be dealing with special relativity in this chapter, it is clear that we will be addressing the issue of the nature of Minkowski spacetime. But any conclusions on the nature of Minkowski spacetime will be valid for any relativistic spacetime, since the kinematic consequences of special relativity that will be examined here are valid in general relativity as well.

that the experimental evidence which confirms them would itself be impossible if the world were three-dimensional.

5.1 Has Special Relativity Posed the Greatest Intellectual Challenge to Humankind?

One of the most difficult problems that science has posed, not only to scientists and philosophers, but to any representatives of humankind, who want their world view to be in accordance with modern science, has come from special relativity. The main question is whether the world is three-dimensional or four-dimensional. It arises from the issue of the ontological status of Minkowski spacetime which leads to a clear dilemma: Minkowski spacetime should be regarded either as nothing more than a mathematical space which represents an evolving-in-time three-dimensional world (the present) or as a mathematical model of a timelessly existing four-dimensional world with time entirely given as the fourth dimension.

The implications of a four-dimensional world for a number of fundamental issues such as temporal becoming, flow of time, free will, and even consciousness are profound – in such a world (often called the block universe) the whole histories in time of all physical objects are given as completed four-dimensional entities (the objects' worldtubes) since all moments of time are not 'getting actualized' one by one to become the moment 'now', but form the fourth dimension of the world and hence are all given at once. And if temporal becoming and flow of time are understood in the traditional way – as involving three-dimensional objects and a three-dimensional world that endure through time – there is no becoming, no flow of time, and no free will in a four-dimensional world. It is these implications of relativity that have posed perhaps the greatest intellectual challenge humankind has ever faced.

With the stakes at the highest level, the reaction of physicists and philosophers to that challenge is quite different. The most frequent answer to the question: "What is the dimensionality of the world according to relativity?" given by physicists and especially relativists is: "Of course, the world is four-dimensional." But when asked about the implications of such a four-dimensional world (not a four-dimensional *mathematical* space), they start to realize that relativity poses serious interpretive problems. Philosophers of science do better, perhaps because one of their *raisons d'être* is precisely the interpretation of scientific theories. Unfortunately, a pattern is easily detected – some philosophers of science who write on issues related to the ontology of spacetime regard the block universe view as undoubtedly wrong and believe that some kind of objective becoming and time flow must exist. The presumption that the world cannot be four-dimensional is sometimes considered so self-evident that any attempts to question it are virtually reprimanded. In 1991 Stein [23] criticized the Rietdijk–Putnam–Maxwell argument [24–26] according to which special relativity implies that reality is a four-dimensional world. Stein argued that their use of the concept 'distant present events' is a fallacy [23, p. 152]: "The fact that such a fallacy (again, if I am right) not only persists among some writers, but is allowed by referees to find continued publication, is a phenomenon that itself calls for reflection." Leaving aside the question of whether Stein is justified in making such a remark, it will be discussed in the next section whether he is right to object to the use of 'distant present events' in the framework of special relativity.

5.2 Relativity and Dimensionality of the World

An attempt to avoid the challenges posed by the interpretation of Minkowski spacetime is to question the very existence of a three-dimensional/four-dimensional dilemma. This was implicitly done by Stein [23, 27] in his criticism of the Rietdijk–Putnam–Maxwell argument which makes use of that dilemma. He argued that Rietdijk, Putnam, and Maxwell were wrong since they incorrectly used the concept of distant present events which is based on the pre-relativistic division of events into past, present, and future [23, p. 159]: "In the theory of relativity the only reasonable notion of 'present to a space-time point' is that of the mere identity relation: present to a given point is that point alone – literally 'here-now'." It has not been pointed out so far that Stein's criticism of the concept of distant present events in relativity is directed rather against presentism and the three-dimensionality of the world, and not against the Rietdijk–Putnam–Maxwell argument. This becomes evident when one takes into account the fact that the only way to define the present (the three-dimensional world existing at the moment 'now') is in terms of the pre-relativistic division of events. What Rietdijk and Putnam showed was that the common view on the world (presentism) is incorrect when relativity of simultaneity is taken into account and therefore reality should be regarded as a four-dimensional world. Stein did exactly the same thing with respect to the first half of their argument – he argued that one could not talk about distant present events in relativity, which meant that the pre-relativistic division of events into past, present, and future could not be used, and therefore an observer could not define his present (or a three-dimensional world) in the framework of relativity. It appears completely unrealistic to assume that Stein would advocate a view according to which one can regard a single event – the event 'here-now' – as the only real one, since such a view clearly amounts to event solipsism.

The fact that presentism is based on the pre-relativistic division of events does not yet constitute a clear contradiction with relativity for two reasons. First, presentists can argue that relativistic causality (reflecting the existence of an upper limit for the speed of physical interactions) does not appear to exclude the possibility that reality can be a three-dimensional world. They might say that the ordinary three-dimensional world is indeed based on the pre-relativistic division of events into past, present, and future, but this does not mean that the three-dimensional objects existing at the present moment of an observer can interact at that moment. And, obviously, the very existence of objects does not depend on whether or not they interact (at the moment they exist). Presentists could point out that every object lies outside of the light cones of the other objects of the three-dimensional world of a given observer. Such a model, in a different context, was discussed by Dieks [28]. Second, as we will see below, the relativity of simultaneity, length contraction, and time dilation all imply pre-relativistic division of events when the issue of existence of the objects involved in these effects is explicitly raised.

In a recent paper McCall and Lowe [29] explicitly questioned the three-dimensional/four-dimensional dilemma. Their main argument is based on the fact that "the three-dimensional and the four-dimensional descriptions of the world are equivalent" and therefore "it is not a question of one being true and the other false" [29, p. 114]. They argue that the equivalence of the three-dimensional and four-dimensional descriptions of the world is an indication that the debate over the three-dimensional/four-dimensional ontologies does not reflect a real problem. Although the equivalence of the two descriptions is not questionable, the equivalence between the three-dimensional and four-dimensional ontologies is. The dimensionality of a physical entity is considered to be one of the basic 'attributes' that determine its very nature. It is like existence – just as an entity cannot be both existent and non-existent, it cannot be both three-dimensional and four-dimensional. As the accepted view is that dimensionality is an intrinsic feature of the world, it follows that the world is either three-dimensional or four-dimensional and any claim that there is no three-dimensional/four-dimensional issue should obviously explain why the accepted view of dimensionality is wrong.

Any attempt to avoid the challenges from the interpretation of Minkowski spacetime by defending the traditional view of presentism or three-dimensionalism is doomed to failure since this view leads to an immediate contradiction with relativity. Presentism is a pre-relativistic view of reality since it is based on absolute simultaneity – the present (the ordinary three-dimensional world) is defined as everything that exists *simultaneously* at the moment 'now'. As presentism is defined in terms of simultaneity, the relativity of simultaneity has an immediate impact on this view: having different sets of simultaneous events, two observers in relative motion have different presents and therefore different three-dimensional worlds. This demonstrates that the traditional presentism contradicts relativity of simultaneity: if it were only the present that existed, then all observers in relative motion would have a *common* present and therefore a common set of simultaneous events; hence relativity of simultaneity would be impossible. However, to say that "anyone who takes relativity seriously cannot take presentism seriously" [30, p. 225] (see also [31, 33, 34] and [35, Sect. 1.4]) is perhaps a little too quick off the mark, although they are ultimately right, as we will see later. Traditional presentism directly contradicts relativity *only* if the existence of the world and the physical objects is considered *absolute* (observer- or frame-independent). With this in mind it seems that there are only two possible ways to avoid the contradiction with the relativity of simultaneity:

- Preserving the absoluteness of existence but giving up the three-dimensionality of the world. In this case the world, whose existence is *absolute*, is four-dimensional and is represented by Minkowski spacetime. Two observers in relative motion, using the ordinary three-dimensional language, will regard two different three-dimensional 'slices' of the absolutely existing spacetime as their presents.
- Preserving the three-dimensionality of the world but giving up the absoluteness of existence. In this case the world is a three-dimensional one whose existence is *relativized* – observer- or frame-dependent. Two observers in relative motion will have different three-dimensional worlds (presents) but each of them will claim that only his own three-dimensional world exists.

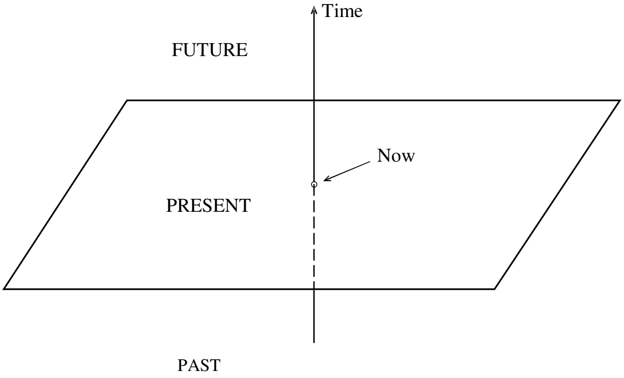


Fig. 5.1. According to presentists, only the present – the three-dimensional world at the moment 'now' – exists

It should be specifically stressed that the relativized existence of an object does not mean that two observers in relative motion describe the existence of the *same* three-dimensional object relative to their own frames of reference. It means, as we will see later, that the observers have different three-dimensional objects, and for this reason the object's existence is ontologically relativized.

Let me explain this in more detail. For this purpose, and for the subsequent analysis of relativistic effects, it will be essential to explicitly address the question: What is the dimensionality of the world at the macroscopic level? (or: What is the dimensionality of the world according to relativity?).

Addressing the question of the dimensionality of the world and physical objects enables us to approach the issue of the existence of the past and the future in the presentist view from a different angle. On the presentist view, to say that the past and the future do not exist appears to amount to a contradiction in terms: since what exists does so only at the moment 'now', the non-existence of the past and the future appears to make sense only if they do not exist in the present. However, when the question of the dimensionality of the world is asked explicitly, things look different. Since on the presentist view, 'exists' and 'exists now' are equivalent, the only reality is the present – the three-dimensional world consisting of all three-dimensional objects, fields, and three-dimensional space that exist *simultaneously* at the moment 'now' (Fig. 5.1). The past is the set of previous *states* of the three-dimensional world, whereas the future is the set of the three-dimensional world's forthcoming *states*. It is for this reason that the

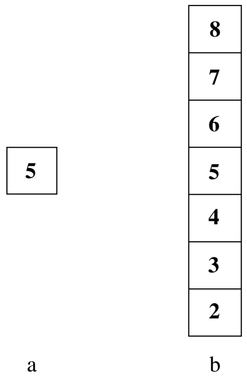


Fig. 5.2. (**a**) According to the three-dimensionalist view, a digital clock exists only at the present moment. (**b**) On the four-dimensionalist view, the digital clock exists at all moments of its history as a four-dimensional object called the clock's worldtube

past and the future do not exist on the presentist view – they are merely *states* of the three-dimensional world which exists solely at the present moment. If the world existed at the other moments too, it would not be a three-dimensional, but a four-dimensional world.

This can be demonstrated by considering how three-dimensionalists (presentists) and four-dimensionalists think of a physical object. As discussed in Sect. 4.1, according to the ordinary (three-dimensionalist) view, a digital clock exists only at its constantly changing present moment indicated by the 5th second on its screen, i.e., the clock exists at the 5th second of its proper time; recall that the proper time of an object is measured by a clock located at the point where the object is (Fig. 5.2a). The presentists contend that the clock exists neither in its past nor in its future. It retains its identity as a three-dimensional object – at all moments of its history it is the *same* three-dimensional clock.

Four-dimensionalists believe that what is real is the worldtube of the clock – a four-dimensional object representing the clock at all moments of its history (Fig. 5.2b). At every different moment, the clock is a *different* three-dimensional object (different 'slice' of the clock's worldtube) which means that it does not preserve its identity as a three-dimensional object. However, what makes the clock the same clock is the fact that it retains its identity as a four-dimensional object – it is a continuous four-dimensional object which is not objectively divided into three-dimensional 'slices' (or three-dimensional cross-sections). We talk of three-dimensional 'slices' in the framework of relativity only when we use our everyday three-dimensional language, which reflects the fact that we perceive three-dimensional images of the physical objects. Figure 5.2 demonstrates why a three-dimensional object can exist only at one moment (at the present moment of its proper time); if an object exists at more than one moment, then that object is four-dimensional[2] – like the worldtube of the clock in Fig. 5.2b. Figures 5.2a and b represent the three-dimensionalist and four-dimensionalist views on what a clock is. The essential question is whether it can be shown that one of these alternative views contradicts relativity.

Now we are in a position to see why presentism can be reconciled with relativity only if the existence of the world and physical objects is ontologically relativized. Consider two observers A and B in relative motion, whose worldtubes are depicted in Fig. 5.3. Two clocks C_1 and C_2 at rest in observer A's frame of reference are also represented by their worldtubes. Using Fig. 5.3 we will try to determine whether the theory of relativity (more specifically, the relativity of simultaneity) can help us determine which view – three-dimensionalist (Fig. 5.2a) or four-dimensionalist (Fig. 5.2b) – is the correct one.

At the moment the observers meet at event M, they set their own clocks (located where the observers are) at the 5th second of their proper times: $t_A = t_B = 5$ s. According to special relativity, A and B have different sets of simultaneous events. Now let us see what the physical meaning of the relativity of simultaneity is by explicitly asking the questions of the existence and dimensionality of the two clocks. On the presentist view (recognizing only the existence of three-dimensional objects), each clock exists only at the moment 'now' of its proper time, as shown in Fig. 5.2a. According to observer A, both clocks exist at the 5th second of the coordinate time measured in A's reference frame. As in an inertial reference frame, the coordinate (global) time coincides with the proper times of all objects at rest in that frame, it follows that A comes to the conclusion that C_1 and C_2 both exist at the 5th second of their proper times. In fact, for presentists all objects existing in A's present (no matter whether they are at rest in A's frame or in motion relative to A) must exist at the present moments of their proper times, since an object exists only at the moment 'now' of its proper time according to the presentist view. This means that

[2] Figure 5.2 is not very rigorous in one respect – the screens of the clocks are extended in the fourth dimension; three-dimensional clocks should be represented by horizontal lines only.

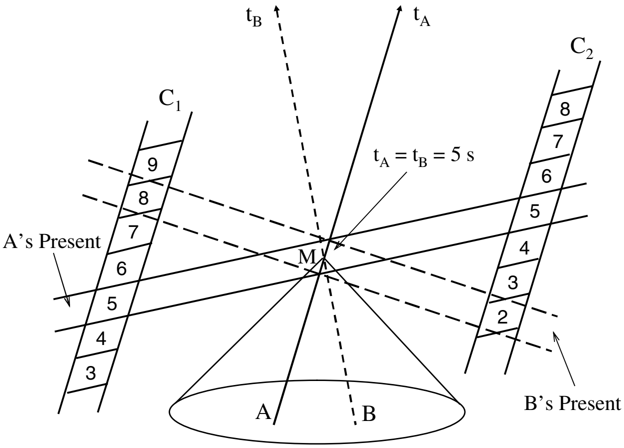


Fig. 5.3. Two clocks C_1 and C_2 represented by their worldtubes are at rest with respect to observer A. The relativity of simultaneity implies that observers A and B, who are in relative motion, have different pairs of three-dimensional clocks which exist in their presents. Each of the clocks must exist at two moments of its history in order for relativity of simultaneity to be possible. For instance, clock C_1 exists at the 5th second of its proper time for A and at the 8th second of its proper time for B

the present moments of the proper times of all objects existing in A's present must be simultaneous for A.

What is simultaneous for A, however, is not simultaneous for B. As shown in Fig. 5.3, what is simultaneous for B at the 5th second of B's time (when B meets A at M) is clock C_1 existing at the 8th second of its proper time and clock C_2 existing at the 2nd second of its proper time. Therefore, for B the moment 'now' of the proper time of C_1 is the 8th second of its proper time, whereas the present moment of C_2 is the 2nd second of its proper time. So, when A and B meet at M, they will disagree on which is the present moment of each of the clocks and on what exists for them at the moment of meeting (at the 5th second of A's time and the 5th second of B's time): for A each of the two clocks exists at the 5th second of its proper time (at its 'now' according to A), whereas for B clock C_1 exists at the 8th second of its proper time (at its 'now' according to B) and clock C_2 exists at the 2nd second of its proper time (at its 'now' according to B). Therefore relativity of simultaneity is possible in the framework of the presentist view if *different* pairs of clocks exist for A and B at M (and at any other moment of A's and B's times while the two observers are in relative motion). At event M, one pair of clocks exists simultaneously with M according to A and another pair exists simultaneously with M according to B. This is only possible if each of the two clocks exists as a different three-dimensional object at every different moment of its proper time. For instance, when A and B meet at M, C_1 exists as a *different* three-dimensional object for A and B – one existing at the 5th second of the proper time of C_1 and belonging to A's present, the other – existing at the 8th second of the proper time of C_1 and belonging to B's present. (This conclusion holds at every moment of A's and B's times while they are in relative motion.)

The immediate consequence of the relativity of simultaneity (as depicted in Fig. 5.3) that a given object exists as a different three-dimensional object at every different moment of its history seems to contradict the presentist view according to which a physical body retains its identity as a three-dimensional object through time (as shown in Fig. 5.2a). However, to see the contradiction between presentism and relativity, we need to clarify what we mean by saying 'an object exists for an observer'. In other words, we must explicitly state whether the existence of the object is absolute (observer- or frame-independent) or relative (observer- or frame-dependent).

If existence is regarded as absolute, it is clear from the discussion of relativity of simultaneity why the traditional presentism (assuming that an object exists only at its present moment) directly contradicts relativity – if the clocks existed at the 5th seconds of their proper times, all observers in relative motion should acknowledge this fact and therefore the *same* pair of three-dimensional clocks (the clocks existing at their 5th seconds) would exist simultaneously for A and B at M. Obviously this would mean that simultaneity would be absolute.

If presentists regard the clocks' existence as absolute but, in view of the relativity of simultaneity, agree that at M different pairs of three-dimensional clocks exist for A and B, they will inevitably arrive at the four-dimensionalist view: each of the clocks C_1 and C_2 will exist at *two* moments of its proper time, and this is only possible if the clocks are four-dimensional objects, as depicted in Fig. 5.2b. The only way for presentism to avoid direct contradiction with relativity is to regard the existence of the three-dimensional clocks as relativized. In the relativized version of presentism, the observers disagree on what exists – for observer A, clock C_1 does not exist at its 8th second since it lies in A's future; C_2 does not exist at its 2nd second either because it already existed at that moment three seconds before the meeting and is therefore in A's past. Observer B denies the existence of C_1

and C_2 at their 5th second since C_1 at its 5th second is in B's past, whereas C_2 at the 5th second of its proper time is in B's future. All this means that every observer recognizes only the existence of the pair of three-dimensional clocks which belongs to his present but denies the existence of the pair of three-dimensional clocks which is part of the other observer's present.

In my view, the concept of existence employed by a relativized version of presentism is so twisted that Nature is unlikely to be impressed by this pushing of the human imagination to such an extreme that allows observer A to claim that C_1 at its 8th second does not exist for him but exists for B. However, even the relativized version of presentism cannot deny the consequence of the relativity of simultaneity that *different* pairs of clocks exist for A and B, which means that every clock exists as a *different* three-dimensional object at the different moments of its proper time. At the same event M, observers A and B claim that clock C_1 exists as *two* different three-dimensional objects for them: C_1 existing at its 5th second is simultaneous with M according to A (and therefore exists for A at the moment of the meeting $t_A = 5$ s), whereas C_1 existing at its 8th second is simultaneous with M according to B (and therefore exists for B at the moment of the meeting $t_B = 5$ s). Observer A acknowledges the existence of the clock C_1 at its 5th second when he meets B at M, but knows (due to the relativity of simultaneity) that C_1 also exists at its 8th second as a different three-dimensional object for B at the moment of the meeting. B is in the same situation – he acknowledges the existence of the clock C_1 at its 8th second but knows that C_1 also exists at its 5th second as a different three-dimensional object for A. This shows why the price presentism should pay to avoid a direct contradiction with relativity is an ontological relativization of existence.

The consequence of the relativity of simultaneity that a given object exists as a different three-dimensional object at every moment of its history is fully consistent with the definition of an event in relativity (discussed in Chap. 4): an object (or a field point, or a point in space) at a given moment of its proper time. In Chap. 4 we have seen that in relativity we cannot talk about *different* events that happen with the *same* three-dimensional object since such a statement would be a contradiction in terms. To see this better consider again clock C_1. According to A and B two different events associated with C_1 are simultaneous with M – the event 'C_1 existing at its 5th second' is simultaneous with M according to A, whereas the event 'C_1 existing at its 8th second' is simultaneous with M according to B. However, these two different events are two different three-dimensional objects – the clock C_1 at its 5th and at its 8th seconds. If C_1 did not exist as *two* different three-dimensional objects at its 5th and 8th seconds, no relativity of simultaneity would be possible no matter whether existence is considered absolute or relative: if existence is absolute, C_1, being a *single* three-dimensional object existing solely at its present moment, exists at either its 5th or its 8th second for *both* A and B; if existence is relative, C_1 exists at its 5th second for A, but A knows that in order for the relativity of simultaneity to be possible, C_1 must exist[3] at its 8th second as a *different* three-dimensional object for B at the moment of the meeting $t_B = 5$ s (otherwise relativity of simultaneity is impossible).

The proper understanding of what an event is in relativity is crucial since one might be tempted to think that the *same* three-dimensional object exists for two observers in relative motion (which appears more than obvious[4] according to our everyday experience!), but *different* events of that object are *perceived* by the observers. The impossibility for one pair of three-dimensional clocks to exist for both observers A and B but the readings of those clocks to be different for the observers is evident from the fact that what is simultaneous for the observers is not what they *see* at their present moments, but what exists simultaneously at their moments 'now'. At M both observers will see precisely the *same* thing – the past light cone at event M (Fig. 5.3). For this reason the fact that two observers in relative motion have different sets of simultaneous events does mean that the observers have *different* sets of three-dimensional objects that exist simultaneously at the observers' present moments, in full agreement with the definition of an event.

Four-dimensionalists have no problem explaining the relativity of simultaneity. They regard the worldtubes of the clocks as real four-dimensional objects, which naturally explains why different pairs of three-dimensional clocks exist for the observers A and B at M. In terms of our everyday three-dimensional language, the observers' presents 'cut off' different pairs of three-dimensional 'slices' from the clocks' worldtubes. However, the worldtubes of the clocks are not objectively divided into three-dimensional 'slices', which shows that, on

[3] As this example shows, an analysis of the concept of relativized existence in the framework of special relativity reveals the meaninglessness of this concept.

[4] It is really the same object that exists for both observers, but not the same *three-dimensional* object; it is the same *four-dimensional* object – the worldtube of the physical object under consideration – that exists for both observers.

the four-dimensionalist view, the concept of a three-dimensional clock is just a *description* and does not have any objective meaning. The analysis of relativity of simultaneity supports the four-dimensionalist view but does not provide a decisive argument against the relativized version of presentism. We have seen, however, that the relativity of simultaneity forces presentists to admit that a physical object must exist as different three-dimensional objects at the different moments of its proper time, which means that the presentist view of a physical object (depicted in Fig. 5.2a) must be relativized to avoid a contradiction with relativity.

It should be noted that when we consider the *existence* of the clocks involved in the relativity of simultaneity as shown in Fig. 5.3, it becomes clear that this effect is formulated in terms of the *pre-relativistic* division of events which is applied to each of the observers. And indeed if, at event M, the observers A and B in Fig. 5.3 ask at what moment of its proper time each of the clocks exists, the only answer that is consistent with relativity is dictated by the relativity of simultaneity: for observer A, both clocks exist at the 5th second of their proper times; for observer B clock C_1 exists at the 8th second of its proper time, whereas clock C_2 exists at the 2nd second of its proper time. Observer A would not say that C_1 and C_2 exist (for A) at the 8th and 2nd seconds of their proper times, respectively, since these two events are not simultaneous with the event of the meeting M. Therefore, when the existence of C_1 and C_2 is considered, relativity of simultaneity makes sense only if each of the observers applies the pre-relativistic division of events into past, present (existing), and future. That is why the presentness (existence) of the remote clocks for the observers A and B is crucial for them if the relativity of simultaneity is to have any meaning – if A and B insisted, when they met at the event M, that only M (the event ‘here-now’) were present and therefore real for them, they would not be able to say anything about the existence of the remote clocks C_1 and C_2 and would not be able to make any statement about what existed simultaneously for them at M. This shows that Stein was not right when he wrote [27, p. 16], [23, p. 155]: “The fact that there is no experience of the presentness of remote events was one of Einstein’s basic starting points.” Strictly speaking, Stein is right, since the concept of present events is indeed incompatible with the concept of spacetime, but that incompatibility implies four-dimensionalism, which Stein rejects. The employment of a pre-relativistic division of events by each of two observers in relative motion when they determine what is simultaneous for them is not surprising since, in its original 1905 formulation, special relativity was implicitly based on the existing world view that it is only the present that exists. In Minkowski's formulation of special relativity, the relativity of simultaneity makes sense only when two observers in relative motion describe the *indivisible* spacetime in terms of our everyday three-dimensional language, which is the basis for the presentist view.

Two other relativistic effects – length contraction and time dilation – originate from the relativity of simultaneity, i.e., from the fact that, having different sets of simultaneous events, the observers in relative motion have different three-dimensional spaces (i.e., different presents according to the presentist view). As relativity of simultaneity is behind these effects, their analysis also demonstrates that presentism can avoid an immediate contradiction with relativity only if the existence of the physical objects is ontologically relativized.

5.3 Length Contraction

Consider two observers A and B in relative motion. A rod of proper length $l_A = Q_A - P_A$ is at rest with respect to observer A and its worldtube is depicted in Fig. 5.4. The middle point of the rod coincides with the origin of reference frame S_A in which observer A is at rest. When the observers meet at event M, they measure the length of the rod. As we have seen in Chap. 4, the radical research team concluded that, if spacetime is real, the worldtube of the rod is a real four-dimensional object and the instantaneous three-dimensional spaces of A and B intersect the rod's worldtube at two different places. This means the observers will determine that the rod has different lengths in the reference frames S_A and S_B, and that its proper length l_A measured in its rest frame S_A is the longest. Therefore when observer B measures the length of the rod, he will find a shorter length $l_B < l_A$ ($l_B = Q_B - P_B$).

Now we know that special relativity predicts the same effect, and the natural question is: Is the spacetime explanation of the length contraction given by the radical research team the only possible explanation? At the turn of the twentieth century, there were attempts to explain it in terms of forces acting between the atoms of the contracted body. However, any attempt to explain the length contraction in terms of a real deformation involving forces faces three insurmountable problems:

- An explanation of length contraction in terms of deformation implies that the relativistically contracted object is the *same* three-

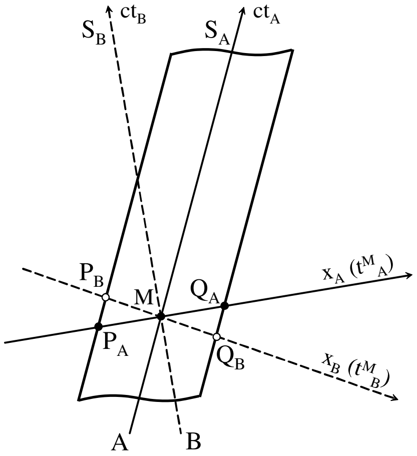


Fig. 5.4. A rod of proper length l_A at rest in observer A's frame is represented by its worldtube. The length of the rod for observer B, who moves relative to A, is relativistically contracted

dimensional object. However, as we have seen in Chap. 4, the assumption that two observers in relative motion measure the *same* three-dimensional object contradicts the relativity of simultaneity; we will return to this point shortly.

- If an object is deformed, it is *objectively* deformed; that is, it is deformed for everyone and not only for some specific observers. In the case of length contraction, however, it is clear that the rod by definition is not deformed for the A-observer; it is contracted only for the B-observer. Due to the fact that this effect is relative or reciprocal, if B has the same type of rod, then A will find that B's rod is shortened and therefore deformed (but for B the rod does not suffer any deformation). A deformation of the rod is described by a stress tensor. But if a tensor is zero in one coordinate system it is zero in all coordinate systems. Since the rod is not deformed for A, the stress tensor in A's coordinate system is zero. Therefore it cannot be different from zero in B's coordinate system, which means that the rod cannot be deformed for B.
- What definitely shows that the force explanation of the length contraction is wrong is that it cannot explain the contraction of *space* (where there are no atoms and no forces that can cause its deformation). In special relativity everything that is in relative motion contracts, including space. For instance, the muon experiment [36] cannot be explained if it is assumed that space does not contract.

> Muons are unstable particles which are produced by cosmic rays in the upper layers of the Earth's atmosphere. As muons are short-lived particles, they should not be able to reach sea level even if they travelled at the speed of light. However, a lot of muons are observed at sea level, and this can only be explained by the time dilation effect – in the Earth's reference frame their half-life is longer than their proper time; so they live longer and can arrive at the surface of the Earth (sea level) before decaying. However, in the muon's reference frame, the muon also reaches sea level since a *point* event (e.g., a muon arriving at the Earth's surface) is the *same* event in all reference frames. This can be explained only if *space* itself contracts.[5]

The true explanation of the length contraction effect is that the three-dimensional spaces of A and B intersect the rod's worldtube at different angles [9], [37, p. 70] – A's three-dimensional space 'cuts' the rod's worldtube in the three-dimensional cross section P_AQ_A, whereas B's three-dimensional space intersects the rod's worldtube in the cross section P_BQ_B and due to the different angles of the two cross-sections $l_B < l_A$.

Many regard the spacetime diagram in Fig. 5.4 as nothing more than a mere graphical representation that should not be taken too seriously. Let us see whether this is really the case. The first thing

[5] Not taking this into account may lead to confusion. J. Bell [38] discussed a thought experiment which he tried to interpret in terms of length contraction. In this experiment spaceships B and C, which are connected with a fragile thread, accelerate gently in such a way that at every moment they have the same velocity relative to an inertial spaceship A: "[...] and so remain displaced one from the other by a fixed distance." Bell claims that, as the spaceships B and C speed up the thread [38, p. 67] "will become too short, because of its need to Fitzgerald contract, and must finally break. It must break when, at a sufficiently high velocity, the artificial prevention of the natural contraction imposes intolerable stress." An obvious problem with Bell's explanation is his assumption that the space between B and C does not contract, whereas the thread does. Also, as a rule, those who believe that length contraction involves forces do not analyze sufficiently the reciprocity of this effect. Had Bell considered that it was spaceship A that slowly increased its velocity while B and C had not changed their state of motion, he would have realized that if the thread broke when B and C were gently accelerating, it should break when A accelerates as well. (It should be noted that the sole role of acceleration in Bell's argument is to ensure a continued increase of the length contraction [38, p. 78, note 3].) But such a break would be a miracle for B and C since their state of motion had not been changed. Thus taking the reciprocity of the effect into account demonstrates that the thread cannot break due to length contraction and therefore the three-spaceship thought experiment cannot be explained in terms of length contraction.

that is immediately evident is that this effect makes sense only if what every observer measures is a three-dimensional object – the rod is by definition a three-dimensional object of proper length l_A. As shown in Fig. 5.2a, a three-dimensional object exists only at its moment ‘now’; an object that exists at once at more moments of its history is a four-dimensional object as shown in Fig. 5.2b. If the three-dimensional object is *extended* like the rod, then all its parts must exist *simultaneously* at the moment ‘now’ of an observer. Therefore, an extended three-dimensional object is defined by an observer in terms of the *pre-relativistic* division of events – all parts of the rod exist *simultaneously* at the present moment of the observer, which means that they constitute a set of present events (the rod existing at the observer’s moment ‘now’). As the observers A and B have different sets of simultaneous events, it follows that they have *different* three-dimensional objects as their rod, which is by definition *one* object. This paradox disappears if what is depicted in Fig. 5.4 represents the true situation – that the rod is indeed *one* object which, however, is four-dimensional, represented by the rod’s worldtube, and the three-dimensional spaces of A and B ‘cut’ it in two different three-dimensional cross-sections.

In order to illustrate even better why A and B have different three-dimensional rods at any moment of their times while they are in relative motion, consider the following thought experiment. As in Fig. 5.4, a rod is at rest in reference frame S_A (Fig. 5.5). When observers A and B meet at event M, they determine the length of the rod in their reference frames. The meeting happens at the moment t_A^M of A’s time and at t_B^M of B’s time. There are lights incorporated at the end and middle points of the rod. Every instant the color of the lights changes simultaneously in S_A: at the moment t_A^{g} all three lights are green, at $t_A^M = t_A^{\mathrm{r}}$ the lights are red, and at t_A^{b} they are blue.

At the instant of the meeting all lights of the rod are red as determined in S_A. Therefore at that moment, what exists for observer A is the red rod – the three red lights are simultaneous for A at his present moment $t_A^M = t_A^{\mathrm{r}}$. The green rod existed for A one instant *before* the meeting and is in his past while the blue rod will exist one instant *after* the meeting and is in his future. The green and blue rods do not exist for A at $t_A^M = t_A^{\mathrm{r}}$, if one insists that only present objects exist.

As observer B has a different class of simultaneous events, at the moment of the meeting t_B^M the lights of the rod will not all be red for B. He will determine that the rod’s front end point, middle point, and rear end point will be green, red, and blue, respectively. (S_B is moving to the left in Fig. 5.5.) This means that the green–red–blue

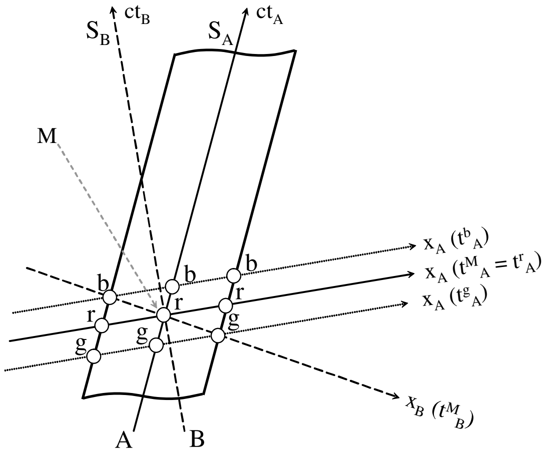


Fig. 5.5. A rod at rest in S_A has lights incorporated at its two end points and at its middle point. The observers A and B, who are in relative motion, meet at event M. In S_A all lights of the rod were green an instant before the meeting with B; they are all red at the moment of the meeting, and their color changes simultaneously in S_A to blue an instant after the meeting

rod, which is present for B, consists of part of the rod that existed in A's past (the front end point with green light), the middle part of the rod (which is also present and therefore exists for A at the moment of the meeting), and part of the rod that will exist in A's future (the rear end point with blue light). As all parts of a three-dimensional object exist *simultaneously* at the present moment of an observer, the three-dimensional rod that exists for B at his present moment t^{M}_{B} is *different* from the three-dimensional rod of A existing at his present moment $t^{M}_{A} = t^{\mathrm{r}}_{A}$. (The event of the meeting M in Fig. 5.4 is the only common present event for both observers.) The rod of each observer is composed of a mixture of parts of the past, present, and future rods of the other observer. Therefore, the conclusion that each of the observers A and B measures a *different* three-dimensional rod is inevitable.

The present analysis shows that what is depicted in Figs. 5.4 and 5.5 is not merely a convenient abstract construct representing the length contraction effect. The very existence of this effect demonstrates that the rod should be a four-dimensional object in order that two observers in relative motion have different three-dimensional rods which are cross-sections of the rod's worldtube.

If presentists regard existence as absolute, they cannot explain this relativistic effect. If the rod existed as a single three-dimensional object as the presentist view holds (and as our everyday experience suggests), then this three-dimensional object would be common to both A and B and no contraction would be possible. The only option for the presentist view is to regard the existence of the three-dimensional rod as ontologically relativized – one three-dimensional rod exists for A and another for B, but A and B recognize only the existence of their own three-dimensional rod.

5.4 Time Dilation

Presentism can avoid an immediate contradiction with the time dilation effect and the experimental evidence supporting it by assuming again that the existence of the clocks involved in this effect is ontologically relativized. Consider two clocks A and B in relative motion, whose worldlines are depicted in Fig. 5.6. The time axes of the two inertial reference frames S_A and S_B in which the clocks are at rest coincide with the worldlines of the clocks. Two instantaneous three-dimensional spaces corresponding to two moments of the time of each frame are depicted in Fig. 5.6; the three-dimensional spaces are represented only by their x axes in the figure. The worldline and the instantaneous three-dimensional spaces of clock B are given with dashed lines.

Let the clocks measure two identical processes that are taking place in the frames S_A and S_B. Every process lasts 5 s as measured in its rest frame. At the moment the clocks meet, two observers at rest in

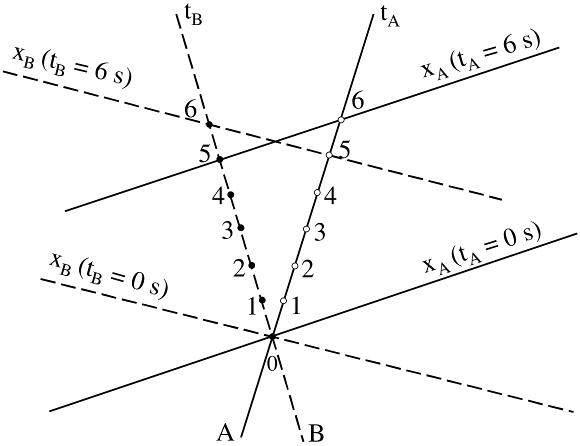


Fig. 5.6. Reciprocal time dilation

S_A and S_B set them to zero and turn on the processes. Let us now see what the duration of the two processes will be as determined by the A- and B-observers.

As the A- and B-clocks measure the proper times in S_A and S_B, respectively, the A-observer will determine that the A-process lasts 5 s in S_A and the B-observer will also find that it takes 5 s for the B-process to finish in S_B. This follows from the relativity principle – the laws of physics are the same in all inertial reference frames. In other words, the proper times of the two observers 'flow' in exactly the same way – one second of the A-clock measured in S_A is equal to one second of the B-clock determined in S_B. In Fig. 5.6, this fact is reflected by the same distance between the marks of two successive seconds on the worldlines of the clocks. For this reason *the worldline of a clock can be regarded as a time ruler*.

Now each of the observers tries to determine the duration of the process taking place in the other frame. In order to measure the end of the B-process, the A-observer should determine which second on the screen of the A-clock is *simultaneous* with the end of the B-process, i.e., with the 5th second on the screen of the B-clock. As seen in Fig. 5.6, the 5th second of the B-clock falls in the instantaneous three-dimensional space $x_A(t_A = 6\text{ s})$ which corresponds to the 6th second of the A-observer's proper time. Since the 5th second of the B-clock (marking the end of the B-process in S_B) is simultaneous with the 6th second of the A-clock, the A-observer concludes that the duration of the B-process as determined in S_A is 6 s. This is what is called the time dilation effect. It is a relative (reciprocal) effect as seen in Fig. 5.6 – the B-observer finds that the A-process lasts 6 s, since the 5th second of the A-clock (the end of the A-process in S_A) falls in the instantaneous three-dimensional space $x_B(t_B = 6\text{ s})$ of the B-observer that corresponds to the 6th second of his proper time.

The time dilation effect looks pretty clear on the spacetime diagram in Fig. 5.6. But that clarity is quite illusory. The most important question that should be asked here is about the existence and dimensionality of the clocks – whether they exist as three-dimensional objects enduring through time or as four-dimensional objects which contain the whole history in time of the ordinary (three-dimensional) clocks and which are represented by the worldlines A and B of the clocks in Fig. 5.6. If the two worldlines of the clocks represent real four-dimensional clocks, the time dilation is indeed clear. As the clocks are in relative motion, their worldlines are not parallel and their three-dimensional spaces do not coincide – they form an angle. As a result, the two instantaneous three-dimensional spaces $x_A(t_A = 0 \text{ s})$ and $x_A(t_A = 6 \text{ s})$ corresponding to the 0th and 6th seconds of the A-observer's proper time 'cut off' different lengths from the two clocks' worldlines – 6 s from A-clock's worldtube and 5 s from B-clock's worldtube.

However, if we believe (on the basis of our everyday experience) that the clocks are three-dimensional objects that evolve in time, then the tough questions start. The first thing that becomes immediately evident is what we have already established in the case of relativity of simultaneity – that the existence of a three-dimensional object cannot be absolute. Consider the time dilation effect determined by the A-observer. In S_A the A-clock with the 6th second on its screen is simultaneous with the B-clock showing the 5th second on its screen. Up to now in discussions of the time dilation effect, close enough attention has not been paid to the fact that the A-observer should implicitly assume that what is real for him at the 6th second of his proper time is everything that exists at that moment, which is his present represented by his instantaneous three-dimensional space $x_A(t_A = 6 \text{ s})$ in Fig. 5.6. That assumption is necessary for the A-observer to conclude that the B-process lasts 6 s in S_A – the B-clock with the 5th second on its screen (indicating the end of the B-process in S_B) comes into existence at the 6th second of the A-observer's proper time. This is the reason why the A-observer regards the B-process as lasting six seconds in S_A. If existence were absolute, both observers should agree that the A-clock existed in its 6th second, whereas the B-clock existed in its 5th second. This obviously contradicts relativity, since the time dilation effect could not be reciprocal if the clocks existed in an absolute manner. It is precisely here that the traditional presentism is manifestly wrong.

The presentist view that the clocks are three-dimensional objects can be preserved only if the clocks' existence is relativized. Then the A-clock in its 6th second and the B-clock in its 5th second will exist in S_A at the 6th second of A's proper time, whereas the A-clock with the 5th second on its screen and the B-clock showing its 6th second will be real in S_B at the 6th second of B's proper time. Here too it becomes evident that *different* pairs of three-dimensional clocks exist for the two observers. Consider clock A. In order for the time dilation to be reciprocal, clock A should exist as two three-dimensional objects – clock A with the 6th second on its screen exists for the A-observer, whereas clock A at its 5th second exists for the B-observer. Therefore different pairs of three-dimensional clocks must exist for the A- and

B-observers. If the observers are presentists and believe only in the existence of three-dimensional objects, then each of them will hold that only his pair of three-dimensional clocks exists and will deny the existence of the other observer's pair of three-dimensional clocks. This, however, is exactly an ontological relativization of the existence of physical objects.

If existence is ontologically relativized, the three-dimensional/four-dimensional dilemma seems to remain in the framework of relativity. At first glance it even appears that special relativity would support such relativization of existence since one is tempted to think that, after having relativized motion and simultaneity, special relativity would require the relativization of existence as well. However, the very idea that the most fundamental 'attribute' of reality – existence – might lose its absolute status and become observer- or frame-dependent in an ontological sense appears disturbing. As Gödel put it [39]: "The concept of existence [...] cannot be relativized without destroying its meaning completely." This can be developed into a strong philosophical argument, but I would prefer to concentrate on arguments demonstrating that special relativity itself rejects the view which regards existence as ontologically relativized.

5.5 Relativization of Existence and the Twin Paradox

We have seen that relativity of simultaneity forces presentists to relativize existence in order to preserve the view of reality as a three-dimensional world. To demonstrate that even a relativized presentism contradicts relativity, let us consider the twin paradox which is an *absolute* effect with no relativity of simultaneity involved.[6] The worldtubes of twins A and B are depicted in Fig. 5.7. Initially A and B are at rest with respect to each other – their worldtubes are parallel before the event D at which twin B departs, and after turning back at event T meets A again at the event M. Twin A's worldtube is a straight line, which means that it is he who does not change his state of motion.

In Euclidean geometry, the straight line is the shortest distance between two points. In Chap. 4 we have seen that in the pseudo-Euclidean geometry of Minkowski spacetime the straight worldline is

[6] Relativity of simultaneity is used when each of the twins describes the rate of the other twin's clock, but the effect itself is absolute and cannot be explained by relativity of simultaneity. As we have seen in Chap. 4, the expression relating the twins' times was obtained by making use of the invariance of the interval.

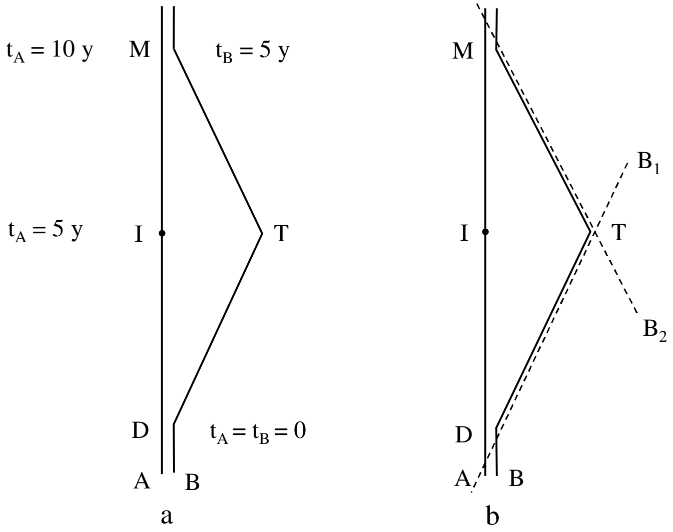


Fig. 5.7. The twin paradox (**a**) and its three-clock version (**b**)

the longest among all worldlines connecting two events. As the proper time of an observer is measured along his worldtube, each of the twins measures his elapsed proper time along his worldtube. The time that has elapsed between events D and M according to twin A is greater than the time as measured by twin B – A's worldtube between D and M is longer than B's worldtube between the same events. (In Fig. 5.7, it is the opposite since the diagram is drawn in the ordinary Euclidean geometry.)

Let us assume that when A and B meet at M, five years have passed for B and ten years for A. Both twins agree that more time has elapsed for A. Thus the time difference between the twins' clocks is an absolute effect – as the twins directly compare their clocks at M, no relativity of simultaneity is involved and no relativization of existence is necessary to explain that difference. What really matters in this relativistic effect is the *direct* comparison of the clocks of A and B when they meet at M; such a comparison involves no relativity of simultaneity. The fact that relativity of simultaneity plays no role in the twin paradox means that it cannot be explained by the reciprocal time dilation since that is based on relativity of simultaneity. Occasionally, physicists and philosophers are tempted to offer a label-placing 'explanation' of this relativistic effect by saying that different time periods have elapsed for the twins because time is frame-dependent in relativity. Obviously, this 'explanation' does not explain anything since the very question is: Why is time frame-dependent in relativity?

The following analysis is intended to demonstrate that the twin paradox effect is possible only in a four-dimensional world, in which the twins' worldtubes are real four-dimensional objects. To see this, let us start from the opposite view – that their worldtubes are not real, and that the twins exist as ordinary three-dimensional objects that evolve as time objectively flows [40]. In such a case both A and B should exist at the event M – otherwise what kind of a meeting would it be if they were not both present there? The only way A and B can explain the time difference of five years is to assume that B's time somehow 'slowed down' during his journey. As the only difference in the states of motion of A and B is the acceleration that B has undergone during his journey, it follows that it should be responsible for the time difference. Also, it is the acceleration that showed the asymmetry between the twins and demonstrated that the twin paradox was not a paradox, but a real effect. However, according to the so-called 'clock hypothesis', the rate of an ideal clock is unaffected by its acceleration [41, p. 164], [42, p. 83], [43, p. 33], [44, p. 55]. And indeed it has been demonstrated that the acceleration does not cause the slowing down of B's time by considering the so-called three-clock version of the twin paradox shown in Fig. 5.7b (see for example [45]). Instead of twin B, who accelerates four times during his journey, consider two clocks B_1 and B_2 which move with constant velocities. At event D the readings of clock B_1 and A's clock are set to zero (when B_1 passes A). When B_1 reaches the turning point at T, it is intercepted by the second clock B_2 and the readings of the two clocks are instantaneously synchronized. The readings of clock B_2 and A's clock are compared at M at the instant B_2 passes A. The calculations show that the difference in the readings of B_2 and A's clock at M will again be five years. As the acceleration does not cause the slowing down of B's time and since no other hypothesis for that slowing down has ever been proposed it appears virtually certain that the flow of B's time is not affected in any way.

The three-clock version of the twin paradox ruled out the acceleration as a possible cause of the difference in the twins' times, but one can still speculate that there might exist some reason for the slowing down of twin B's time. That A's and B's times flow in exactly the same way follows rigorously from the fact that A's and B's clocks measure proper times. To my knowledge the fact that at the event M the twins compare their *proper times*, which according to the relativity principle are subjected to *no* dilation, has been overlooked. As we have seen in Fig. 5.6, the A-process and the B-process take the same amounts of the A- and B-observer's proper times, respectively. What is time dilated is the duration of the B-process as measured by the A-observer and vice versa. As proper times are not relativistically dilated, the proper times of observers in relative motion (existing at their present moments as three-dimensional objects according to the presentist view) must flow *equally*. This means that if the clocks were three-dimensional objects (like the one shown in Fig. 5.2a), in the three clock version of the twin paradox (Fig. 5.7b), where only inertial motion is involved, there would be no time difference when the clocks A and B_2 directly compare their proper times at M. Therefore the twin paradox effect would be impossible.

Let us consider the spacetime diagram in Fig. 5.7a. According to the presentist view, the twins' worldtubes are not real four-dimensional objects – they are rather trajectories along which the twins' three-dimensional bodies evolve in time. As A's and B's proper times objectively flow in the same way, if five years have passed for B (when he exists at event M), five years would have elapsed for A as well, and he would exist at event I. Therefore, in terms of the spacetime diagram in Fig. 5.7a, A and B could not meet at all. The impossibility of the twin paradox effect, if formulated in terms of the presentist view, shows the incorrectness of our initial assumption – that A and B exist only as three-dimensional objects which are subjected to an objective flow of time as required by the relativized version of presentism. The fact that the twin paradox effect and the experiments that confirm it would not be possible if the twins were three-dimensional objects rules out the relativized version of presentism.

The twin paradox is consistently explained if A's and B's worldtubes are real four-dimensional objects; then twin A exists not only at event M (where he meets with B) and event I, but at all events comprising his worldtube. The time difference of five years when the twins 'meet' at M comes from the different lengths of the twins' worldtubes between the events D and M; that is, the different amounts of proper times of the twins between D and M. The four-dimensionalist view offers a natural explanation of why the acceleration does not affect the amount of proper time measured by twin B in Fig. 5.7a: the acceleration which twin B suffers is merely an indication that his worldtube is curved, but this curvature does not affect his proper time since the length of a worldtube does not change if it is curved. The only role of the acceleration in Fig. 5.7a is to show that B's worldtube is curved and that it is a different path from event D to event M which due to the pseudo-Euclidean nature of Minkowski spacetime is shorter than the path along A's worldtube. If B's worldtube is straightened and superimposed on A's worldtube, the length of B's worldtube will be equal to the segment DI of A's worldtube. No matter how mysterious it might look, the twin paradox is merely the triangle inequality in the context of the pseudo-Euclidean geometry of Minkowski spacetime.

5.6 Why Is the Issue of the Nature of Spacetime So Important?

For physicists the answer to this question is twofold:

- in view of the multi-dimensional spaces of modern physics, it appears natural to address the question of the nature of spacetime first,
- if the macroscopic world is indeed four-dimensional, Minkowski's program – physical laws might find their most perfect expression as relations between worldlines – should be pursued more rigorously.

An attempt to follow Minkowski's idea will be made in Chap. 10, where inertia will be regarded as originating from a four-dimensional stress which arises in the deformed worldtube of an accelerated body.

The issue of the ontological status of spacetime is no less important to philosophers of physics, philosophers, and all who want to make certain that their world view does not contradict modern science, since a number of fundamental issues look completely different in a three-dimensional and in a four-dimensional world. Let us briefly discuss the implications of the question of dimensionality of the world for the issues of conventionality of simultaneity, temporal becoming, flow of time, consciousness, and free will.

5.6.1 Conventionality of Simultaneity

Conventionality of simultaneity is possible only in a four-dimensional world. If the world is three-dimensional, simultaneity cannot be conventional. At first glance this is an extremely controversial statement but its correctness becomes evident when the issues of:

- the existence and dimensionality of the world,

and

- conventionality of simultaneity

are analyzed together. As the three-dimensional world coincides with the present (everything that exists *simultaneously* at the moment 'now'), conventionality of simultaneity implies conventionality of what exists, which is clearly unacceptable. Therefore the conventionality thesis turns out to be an argument for the four-dimensionality of the world [46, 47]. And indeed in the Minkowski four-dimensional world, simultaneity is unavoidably conventional – as all events of spacetime are equally existent, it is really a matter of convention which three-dimensional cross-section of spacetime we regard as our three-dimensional world.

Due to the link between conventionality of simultaneity and the dimensionality of the world, any argument for the reality of spacetime is an argument for the conventionality thesis. Any argument in favour of the conventionality of simultaneity is an argument for the four-dimensionality of the world. In a recent paper Dieks [48] pointed out that the conventionality of simultaneity is unavoidable in a rotating reference frame: "Neither the Einstein light signal procedure, nor the slow transport of clocks can be used to establish a global notion of simultaneity on the rotating disc [...]. These non-inertial frames of reference, and general relativistic space-times, seem an arena where the thesis that (global) simultaneity is conventional can be defended without controversy." The fact that this argument is valid only in rotating frames obviously cannot be interpreted to mean that the world is four-dimensional *only* for observers in rotating frames. Those observers do not need the arguments discussed above to conclude that they live in a four-dimensional world. The very fact that simultaneity is conventional in such frames implies the four-dimensionality of the world. Otherwise, if the rotating observers insisted that what exists is the ordinary three-dimensional world, it would be up to their *choice* to decide what would be regarded as *simultaneously* existing at their present moment; that is, what the three-dimensional world is would turn out to be conventional. But if the world is four-dimensional for the rotating observers, it must be four-dimensional for all other observers as well.

5.6.2 Temporal Becoming

Change, passage, and temporal becoming have their ordinary meaning only in a three-dimensional world. Only a three-dimensional body (which preserves its identity in time as a three-dimensional object) can undergo objective change in the sense that it is the *same* three-dimensional object that changes. In a Minkowski world, there is no change since the whole history in time of a three-dimensional body is entirely given as the body's four-dimensional worldtube. As we have seen in Chap. 4 and at the beginning of this chapter, the body's worldtube consists of different three-dimensional objects, since the body exists as *different* three-dimensional objects at the different moments of its history, i.e., as different three-dimensional cross-sections of the body's worldtube. What makes the body the same body is the fact that its worldtube retains its identity as a four-dimensional object in spacetime.

The different regions of the body's worldtube are different but this cannot be interpreted as a relativistic analog of the ordinary concept of change. The reason is that a change means that, when a changed object exists, the object before the change does not exist any more. The change along a worldtube is a completely different kind of change – it is the same type of change we find when we look at different regions of extended three-dimensional bodies (e.g., a pen). The fact that three-dimensional bodies are extended in space, whereas a body's worldtube is extended in time is insignificant because in spacetime spatial and temporal dimensions are equally existent. The spacetime signature $+---$ (or $-+++$) tells us that the nature of the spatial and temporal dimensions is different, but it does not mean that the fourth (time) dimension does not exist as the spatial dimensions do. (If the time dimension were not entirely given like the spatial dimensions, the Minkowski world would not be four-dimensional.)

As the whole histories of all three-dimensional objects are entirely given in a four-dimensional world, this demonstrates that everything *exists* there – there is *no coming into existence*. That is why there is no change, passage, and temporal becoming in a four-dimensional world.

5.6.3 Flow of Time and Consciousness

The concept of time flow has a completely different meaning in three-dimensional and four-dimensional worlds. In a three-dimensional world, we have the ordinary objective and universal flow of time – events are *objectively* divided into past, present (occurring simultaneously at the moment 'now'), and future. In the Minkowski world, all events are *equally* existent and therefore are not objectively divided into past, present, and future. It is usually stated that there is no global 'now' in spacetime. In fact the situation is even worse for the proponents of the view that time flows objectively – the equal existence of the events of spacetime means that there is no local 'now' either. On the objective time flow view, the present moment is privileged (the only moment of time that exists), whereas in spacetime, all events of the worldline of a particle are equally existent and therefore no event is privileged as the particle's 'now'. In such a four-dimensional world, the only meaningful concept of time flow appears to be the one described by Weyl [1]:

> The objective world simply *is*, it does not *happen*. Only to the gaze of my consciousness, crawling upward along the life line of my body, does a certain section of this world come to life as a fleeting image in space which continuously changes in time.

It was Minkowski's four-dimensional formulation of special relativity that marked the first time that the concept of consciousness was needed for the interpretation of a physical theory. The original three-dimensional formulation of special relativity given by Einstein in 1905 did not need the concept of consciousness for its interpretation. This fact demonstrates that the question of the dimensionality of the world is related to the issue of consciousness. One may object that the four-dimensionalist view does not need the concept of consciousness, since we can obtain all results of special relativity without it. It is true that relativity itself as a physical theory does not need that concept. However, special relativity is telling us something very disturbing about the world and we must not only verify the relativistic four-dimensionalist view, but also be prepared to reconcile the reliable pieces of knowledge deduced from our perceptual information with that view, if the world is really four-dimensional. Such a reliable piece of perceptual knowledge is something that no one questions – that we realize ourselves and the world at the constantly changing moment 'now'. Up to now, no one has managed to find a way, which does not involve our consciousness, to reconcile the four-dimensionalist view and the fact that whatever we perceive happens only at the present moment. This seems to indicate that Weyl's proposal holds the greatest promise for a resolution of the apparently insurmountable contradiction between relativity and our experience.

Weyl's view is especially important in demonstrating that those who regard the four-dimensionalist view as obviously wrong stand on shaky ground. Assume, for the sake of argument, that the world is really the way Weyl described it. Your consciousness would crawl upward along the worldtube of your body and read the information from your senses that is stored in your brain, but would incorrectly interpret this information in the sense that a three-dimensional world exists and is constantly changing in time. You would be completely convinced that you were living in a three-dimensional world which is evolving in time. If you read a paper by some scientists who argued that the external world were four-dimensional, you would most probably be quick to declare such a view total nonsense. There would be no way for you to discover that the real world is not three-dimensional if you were building your world view and your philosophical doctrines on the basis of information coming essentially from your senses. How do we know that we are not in the same situation?

Weyl implicitly defined consciousness as an entity which is 'moving' along the worldtubes of our bodies and which makes us aware of ourselves and the external world at one moment. Unfortunately, Weyl's idea, although frequently quoted, did not receive the attention it deserved in the light of the tough problems encountered by the view of objective flow of time. As the view that time objectively and universally flows appears doomed, the alternative view – that the flow of time is mind-dependent[7] – outlined by Weyl should have been examined more rigorously. In order to see the need for such a study, let us briefly discuss the implications of Weyl's idea.

At first glance this idea appears to be self-contradictory since Weyl assumed that consciousness (leaving aside the question of what consciousness itself is) *moves* in Minkowski spacetime where no motion is possible. In fact, there will be a contradiction only if it is assumed that either consciousness 'operates' at the macroscopic level of the world modelled by Minkowski spacetime or Minkowski spacetime is applicable to all levels of reality including the level where consciousness might 'operate' [51]. However, it does not appear realistic to expect that any *macroscopic* concept (such as Minkowski spacetime) will be applicable to *all* levels of reality. At some level lying 'beneath' the macroscopic level of our everyday experience, the properties of the world will inevitably be quite different from what we now know from our macroscopic experience; we have already started to observe such discrepancies at the quantum level. With this in mind it is not unthinkable to expect consciousness to 'operate' at a sub-microscopic level, where the frozenness of the macroscopic world deduced from special relativity does not hold any more.

To demonstrate the most provoking consequences of Weyl's idea, imagine that you concentrate on how you realize yourself. You are

[7] The link between time flow and consciousness – or rather between time and the soul – was discussed by Aristotle [49] and Saint Augustine [50].

aware of your own body only at one constantly changing moment, which we call the moment ‘now’. According to Weyl, the reason for this is that our consciousness crawls upward along the worldtube of our body, realizing at each time only a small region of the worldtube. A consequence from here is that the entity we know almost nothing about – consciousness – is to some extent independent of our body, since our past and future bodies exist just as bodies – our consciousness is not there. The consciousness is always localized in an extremely small area of our worldtube which we perceive as our present body (the duration of that area, i.e., the duration of ‘now’, is still unknown). We cannot assume that our consciousness is spread along the whole worldtube of our body, because such an assumption leads to an obvious contradiction with the fact from our everyday experience – that we realize ourselves only at the moment ‘now’. If all our bodies – past, present, and future – possessed consciousness, then our life would be quite different: we would realize ourselves and the world at *all* moments of our life and there would be no flow of time; we would be in an eternal God-like state.

So, on the four-dimensionalist view, our consciousness does ‘move’ toward the future part of the worldtube of our body, leaving our past bodies consciousnessless and ‘giving life’ to our also consciousnessless future bodies. The natural question that immediately follows from Weyl’s view is what happens to the consciousness when it reaches the end of the worldtube of one’s body. Unfortunately, this question cannot be answered on the basis of the two facts that led Weyl to the mind-dependent view of time flow – that (i) special relativity describes the world as four-dimensional and (ii) we realize the world and ourselves at the present moment.

We are now in a position to complete the explanation of the twin paradox depicted in Fig. 5.7a. The twins exist at all events of their worldtubes, but each of them realizes himself only at a single event when his consciousness reaches and realizes that event. There should be no difference in the mind-dependent flow of time of the twins (the advancement of the consciousness of each of them along his worldtube); at least there is no macroscopic reason, as we have seen above, that can cause any change in the time flow for one of the twins. Then when five years have passed for twin B and his consciousness reaches the event of the meeting M, he will be happy to meet his brother. However, this will be a very strange meeting – twin B will be meeting his brother from his future. As twin A’s consciousness moves in the same way as B’s consciousness, five years have elapsed for A as well and his consciousness realizes event I; so his consciousness is five years behind B's consciousness. When twin A realizes event M, he will be meeting his brother from his past; B's consciousness will be five years ahead.

This provoking explanation of the twin paradox follows unavoidably from Weyl's idea. And that explanation has perhaps an even more provoking implication – can we be certain that some of the people we meet are not just consciousnessless bodies? The reader has perhaps already realized that with this question we have entered the field of the philosophy of mind and more specifically the issue of whether there are other minds.

It appears that Weyl's idea leads to absurd consequences. But nevertheless, it should not be ignored because of that, and indeed should be thoroughly examined, since placing labels has never been a solution to anything, especially to scientific problems.

5.6.4 Free Will

On the presentist view the future is ontologically undetermined, which appears to mean that we possess the free will to be the masters of our own fate. In the Minkowski four-dimensional world, there is no free will, since the entire history of every object is realized and given once and for all as the object's worldtube. Therefore, free will may exist only in a three-dimensional world. Here again the importance of the question of the nature of spacetime is quite evident – whether or not we possess free will depends on the dimensionality of the world.

It may appear shocking that we have no free will if the world is a four-dimensional place in which our life is entirely predetermined. Such a reaction may be a little premature since on the four-dimensionalist view virtually everything in the world looks completely different, including the issue of free will. It is beyond the scope of this book to examine in more detail the impact of the four-dimensionalist view on such concepts as free will and the meaning of life. However, the issues discussed in this chapter provide sufficient information to enable us to reflect on the issues discussed here with an open mind, and this may help us to realize that, even if the world turns out to be four-dimensional and our bodies do not have any free will, life does not necessarily look meaningless.[8]

[8] As another exercise in creative and analytical thinking, assume that the world is indeed four-dimensional. We would certainly be amazed to discover that everything in the world looks completely differently, not necessarily for the worst.

5.7 Summary

By addressing the question of dimensionality of objects involved in relativity of simultaneity, length contraction, and time dilation, it has been shown that the presentist view can avoid a direct contradiction with relativity if the existence of those objects is regarded as ontologically relativized. However, even this relativized version of presentism contradicts relativity, as shown by the analysis of the twin paradox in Sect 5.5. Therefore the only view that is consistent with relativity is four-dimensionalism.

It has also been shown that a number of fundamental issues such as conventionality of simultaneity, change, passage, temporal becoming, flow of time, consciousness, and free will look completely different in three-dimensional and four-dimensional worlds. This means that the question of the nature of spacetime precedes those issues and for this reason should be resolved first.

6 Quantum Mechanics and the Nature of Spacetime

A way to try to avoid the challenges posed by special relativity is to question its validity implicitly.[1] This is usually done by saying that special relativity is not telling the whole story about the world, and for this reason one should not take seriously the arguments for the reality of spacetime based upon it. Then it is pointed out that there are more modern theories, such as general relativity, quantum mechanics, quantum gravity, string theory, etc., in which the question of the nature of spacetime may have a different answer.

Such claims amount to questioning the very validity of special relativity. As we have seen in Chaps. 4 and 5, the arguments supporting the reality of spacetime are consequences of special relativity. So, to say that special relativity is not revealing the whole truth about the part of the world it describes, which in the given context means the whole truth about the nature of spacetime (i.e., the truth about the dimensionality of the world on the macroscopic scale), amounts to questioning the theory itself. One may object that, before doing that, it is more natural to argue that the consequences of special relativity can be explained in the framework of the three-dimensionalist view. That is true, but whenever one resorts to the modern theories argument, it is only when one has failed in every attempt to show that the spacetime explanation of the consequences of special relativity is not the only explanation.

The best way to see why the modern theories argument fails is to recall the analysis carried out in Chap. 5 – that it is not just the consequences of special relativity (as *theoretical* results) that would be impossible if the world were three-dimensional (i.e., if spacetime were

[1] The still existing attempts to question special relativity explicitly will not be discussed here. Unfortunately, their authors have failed to recognize so far one of the most valuable lessons of the history of science – that science never goes backwards. One of the goals of this book is to show that relativity is not only inevitably correct, but will never be proven wrong in its area of applicability, as we will see in the next chapter.

not real); it is the *experimental evidence* itself, which confirms those consequences, that would be impossible if the world were not four-dimensional. So the arguments supporting the reality of spacetime are much stronger, since they are derived directly from the experiments which confirmed the consequences of special relativity.[2] For this reason the only relevant attack against those arguments is to try to *prove* that the experimental evidence confirming the kinematic consequences of special relativity can be *explained* (not merely *described*) in terms of the three-dimensionalist view.

Those who advance the modern theories argument seem to believe that a theory that comes after special relativity may provide a new way to interpret the experimental evidence. We will see in this and the next chapters that such an expectation does not seem to have any justification. Let us start with general relativity since it was the first theory to come after special relativity. An analysis of general relativity similar to the one carried out in Chap. 5 not only does not question the arguments for the reality of spacetime derived from special relativity, but it even provides new arguments. The kinematic consequences of special relativity – relativity of simultaneity, length contraction, time dilation, and the twin paradox – are valid in general relativity as well, and their interpretation is the same. The analysis of the proper general relativistic effects also demonstrates that these effects are manifestations not only of a real four-dimensional world, but of a curved four-dimensional world.[3] For instance the gravitational red shift can be explained only if spacetime is real which makes it possible for two observers at different locations in a gravitational field to have *different* proper times.

As quantum gravity and string theory are not experimentally confirmed theories, let us see whether quantum mechanics can have any effect on the debate over the nature of spacetime.

[2] Recall the beginning of Minkowski's talk in 1908 [9]: "The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength."

[3] Such an analysis shows that the following general argument holds. The very fact that, in general relativity, gravity is a manifestation of the curvature of spacetime shows that general relativity presupposes the reality of spacetime; otherwise how can something non-existing be curved?

6.1 Quantum Mechanical Arguments Against the Reality of Spacetime

The basic idea that quantum mechanics may influence the debate over the nature of spacetime is the following. As special relativity is a classical (non-quantum) theory, the concept of spacetime is sometimes seen as an extreme version of the rigid determinism of Newtonian physics – the histories of all objects are realized in their worldlines, and in this sense are completely determined. Quantum mechanics, on the other hand, is a probabilistic theory – the histories of the quantum objects cannot be predicted with certainty. Therefore, the argument goes, the world is probabilistic and one should not bother about the implications of a classical (non-quantum) theory such as special relativity.

This argument is wrong on several counts. As quantum mechanics does not predict the relativistic effects, it does not contain even the possibility of any new interpretation of the experimental evidence confirming them. The quantum mechanical argument is also based on a misconception concerning spacetime. We have seen that the relativistic effects are possible if spacetime is real, but the reality of the worldlines of objects does *not* imply in any way that the future histories of these objects can be *predicted* with absolute certainty.[4] This shows that the reality of the worldlines of objects and whether or not the prediction of the future histories of these objects are deterministic or probabilistic are two different issues. We will return to this point in the next section.

The irrelevance of the quantum mechanical argument for the debate over the nature of spacetime is perhaps best seen from the fact that you are able to read this text. The equations of motion of quantum mechanics govern the behaviour of quantum objects at the microscopic level, where electrons, protons, neutrons, atoms, etc., live. All these entities act in accordance with the probabilistic laws of quantum mechanics. However, when the quantum mechanical equations of motion are applied at the macroscopic level, according to Bohr's correspondence principle, they should coincide with the classical equations of motion. In the case of spacetime this means that no matter how the constituents of the worldline of a given object behave – deterministically or probabilistically – it will not affect the shape of the

[4] When we solve the classical equations of motion (Newtonian or relativistic), we make use of given initial conditions which have to be determined experimentally. Any uncertainty in the initial conditions leads to greater uncertainties in predicting the future states of the objects.

worldline.[5] To illustrate this situation better, consider the letters of the text you are now reading. Each letter consists of a large number of ink particles. Each ink particle contains billions of electrons, protons, etc., whose behaviour is probabilistic. If the probabilistic nature of quantum mechanics were also manifested at the macroscopic level (the level of the letters on this page), then all letters would also behave in a probabilistic manner – constantly changing their shape and jumping around.

6.2 Is Quantum Mechanical Probability Objective?

In this section we will examine in more detail whether or not the existence of spacetime and the probabilistic nature of quantum mechanics are in conflict. We have seen that there is no conflict at the macroscopic level. However, there are no indications that the dimensionality of the world at the microscopic scale, where quantum mechanics operates, is different from that at the macroscopic level. Therefore spacetime should apply at the level of quantum mechanics as well. This implies that the whole histories in time of the quantum objects are also entirely given, which appears to show that the probabilistic predictions of quantum mechanics unavoidably contradict the four-dimensionalist view.

Such a contradiction, if it existed, would constitute a real crisis in physics. On the one hand, the experimental evidence that confirms the consequences of special relativity can be interpreted only in terms of a real spacetime. On the other hand, however, the probabilistic predictions of quantum mechanics are also experimentally confirmed. As experimental results cannot contradict each other, it appears at first glance that either the spacetime interpretation of the relativistic effects is wrong or the probabilistic interpretation of the experiments confirming quantum mechanics is not true. In cases like this the first thing to do is to examine whether the contradiction leading to that dilemma is real.

[5] That the shape of a worldline cannot change follows independently from the definition of spacetime – time is entirely given as the fourth dimension and therefore no change is possible in spacetime. A worldline might change its shape if there existed a second time, but such a hypothesis should be ruled out at least on the macroscopic scale since, if two observers were at different spacetime distances from a given event, they would *see* different outcomes (versions) of the event. Such a thing has never been observed.

It is tempting to assume that the real behaviour of the quantum objects is deterministic but that our knowledge about them is incomplete, and for this reason we have to use a probabilistic description. This would mean that, as in the classical case, the worldlines of the quantum objects might also be entirely given and therefore the objects' probabilistic description is caused by our incomplete knowledge and does not reflect an objective (ontological) probability. We will see shortly that such an interpretation contradicts the existing experimental evidence.

If probability in quantum mechanics is not epistemological (reflecting our incomplete knowledge) but ontological, the natural question is how objective probability should be understood. It seems it is taken for granted that objective probability means an open future – a future which is not given (as the three-dimensionalist view holds) and therefore the future histories of quantum objects cannot be predicted with certainty. If this were the case, then the contradiction between quantum mechanics and the four-dimensionalist view would be unavoidable. But is this really the case?

That three-dimensionalism (according to which the future is open) does not necessarily imply objective probabilism is best demonstrated by the fact that, on the three-dimensionalist view, classical theories (such as Newtonian mechanics and special and general relativity) are also regarded as *deterministic* (*not* probabilistic) theories. So how should we understand ontological probability? What makes this question especially difficult for some scientists is their strong intuition that there is no room for ontological probability in Nature since she would not 'know' herself if probability were objective. Einstein's famous "God does not play dice" expresses the same attitude toward this understanding of probability.

The issue of the nature of probability would have been an exciting research project for the radical research team. It would be rewarding for the reader to analyze that issue and see what conclusions can be drawn. We will not perform such an analysis here since that alone would take at least a whole chapter. However, we will approach the same issue from a different angle and reach the same unexpected conclusion, namely that the behaviour of an object is deterministic if it exists *continuously* in time, whereas an objectively probabilistic behaviour of a particle implies that the particle's existence is *discontinuous* in time. We will obtain another surprising result – that the existence of objective probability does not entail that Nature does not 'know' herself.

6.3 The Nature of the Quantum Object and the Nature of Spacetime

In 1926 Born gave a probabilistic interpretation of the wave properties of particles – of the electron, for example. He suggested that the wave that is 'attached' to the electron was not a real wave but a wave of probability – the probability of finding an electron at a given point. Since then the standard interpretation of quantum mechanics – called the Copenhagen interpretation – has been based on this probabilistic interpretation of the wave properties of the elementary particles.

The Copenhagen interpretation of quantum mechanics, however, does not answer the question as to what the quantum object is when it is not measured. It declared this type of question to be meaningless, since the only information we have comes from experiment and therefore we cannot ask questions about what we have not measured. But the very fact that an electron exists before being measured implies that it should be somewhere. This kind of argument forced some proponents of the Copenhagen interpretation to make extreme claims, e.g., that the quantum object does not exist during the time when it is not measured. Now no one takes such claims seriously. But the mystery remains. As Feynman put it: "If you think you understand quantum mechanics, you misunderstood it."

To realize the difficulty better, consider an electron in a room. According to quantum mechanics there exists a probability of finding the electron anywhere in the room. If the electron were a point-like (localized) particle, then quantum mechanics would not be an adequate theory since the electron would be at a given place at a given time, but the theory will tell us that it could be found with a given probability at any place in the room. If we assume that the electron is not localized but is some kind of fluid, which occupies the whole room, then why can we not isolate a fraction of its charge? Also, in this case, the electron should collapse *instantaneously* to the point where it is measured, which would contradict relativity. So the electron can be neither a small (point-like) particle nor a kind of fluid which occupies the whole volume where the electron's wave function is different from zero. On the other hand, the electron is always measured as a *localized* particle.

There have been different attempts to resolve the wave–particle duality of the quantum objects leading to the above difficulties. One of them is especially telling. Quantum mechanics predicts that the electron in its ground state in the hydrogen atom can be found with a given probability around the nucleus (the proton). If the electron were regarded as a particle, i.e., as localized somewhere above the proton, then the hydrogen atom should possess a dipole moment in its ground state. Both quantum mechanics and experiment show that this is not the case. One may picture the electron as so rapidly orbiting the proton that what is experimentally measured is the average value of the dipole moment over the measurement time. And since there is a spherical symmetry in the ground state, all dipole moments cancel out exactly – the average value is zero. To verify this hypothesis Madelung[6] calculated the orbital velocity of the electron that would ensure that all dipole moments during the measurement cancel out. It turned out that the electron orbital velocity should be several orders of magnitude greater than the velocity of light. This shows that the electron charge should be somehow *actually* distributed around the proton. In other words, *the electron should exist everywhere where the probability of finding it is different from zero*. Such a requirement resolves the problem about the lack of a dipole moment of the hydrogen atom, but the fact that a fraction of an electron has never been isolated remains unexplained. This is a typical quantum mechanical paradox – an electron is always measured as a localized entity, but before its detection, it occupies the whole volume where its wave function is not zero; the spread of the electron in the volume, however, cannot be detected.[7]

Let us now see whether we can employ some of the epistemological lessons drawn from our discussions of different paradoxical situations in Chap. 3. The main paradox in quantum mechanics is the wave–particle duality of the quantum object – the quantum object is always measured as a localized particle, but when no experiment is carried out it occupies the whole volume where the probability of finding it is different from zero. Obviously, what we have to do is analyze the meaning of the two apparently mutually exclusive states of the quantum object – being localized and being everywhere in the volume

[6] In May 1989 I had to leave Bulgaria on a 24-hour notice and my archive was lost. Since then I have not been able to locate the reference to Madelung's paper.

[7] It will be a natural first reaction to say that if something cannot be detected, it is almost certain that it does not exist – the radical research team ruled out the existence of absolute uniform motion precisely for this reason. So should we conclude that an electron does not exist in the volume before being measured? Situations like this are different. The existence of the hydrogen electron around the nucleus is *indirectly* demonstrated by the fact that the hydrogen atom does not posses a dipole moment in its ground state. Another example is the existence of virtual particles. They can never be directly detected but have indirect manifestations.

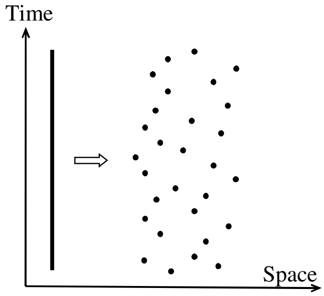


Fig. 6.1. The idea of 4-atomism – an electron is represented not by its worldline (the *solid line*), but by the points constituting the worldline

where it is trapped. Our common sense tells us that the same object cannot behave in both ways. I believe we have become accustomed to such paradoxical situations and already know the general way out: the paradox is caused by an implicit assumption which is wrong.

There are many similar cases in the history of science. Perhaps the most appropriate for our analysis is Zeno's paradox "The Dichotomy" – a finite distance cannot be travelled since its half should be first travelled, then the half of the other half and so on to infinity. Zeno arrived at this paradox since he implicitly assumed that space was infinitely divisible, but time was not. Here we see virtually the same situation. The quantum object is always measured as an entity *localized in space*. But what about time? We have been implicitly assuming that the quantum object exists *continuously* in time; so it is not *localized* in time. Once we have realized that implicit assumption we can understand why we have reached the main quantum paradox – that an electron, for instance, is always detected as a localized object, but is everywhere in a given region when not measured.

For this purpose let us now first see why a particle which exists *continuously* in time cannot be everywhere in a volume when not measured. To do this assume that the quantum object exists discontinuously in time. Such an assumption amounts to bringing the idea of atomism to its logical completion – discreteness not only in space but in time as well: 4-atomism[8] [52,53]. In this case the electron will be represented not by its worldline (as deterministically described in special relativity) but by a set of points (quantons) in spacetime – the 'disintegrated' worldline of the electron (Fig. 6.1). The Compton frequency

[8] This radical idea was developed in the early 1980s by Anastassov of Sofia University. Unfortunately, it has remained unnoticed so far.

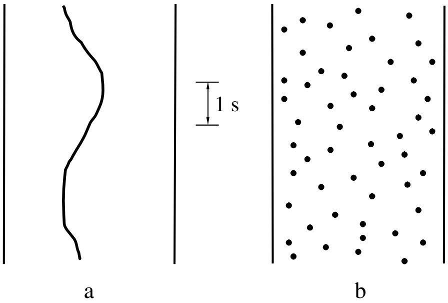


Fig. 6.2. (a) A point-like electron that exists continuously in time is trapped in a room which is represented by the worldlines of two of its walls. During a given period of time (say, 1 s) such an electron, when not measured, cannot be everywhere in the room, where the probability of finding it is different from zero. (**b**) During the same period of time an electron which exists discontinuously in time occupies the whole volume of the room

of the electron can be interpreted to mean that for one second an electron is represented by 10^{20} quantons. According to the 4-atomistic hypothesis the quantons constituting the electron will be scattered all over the spacetime region in which the wave function of the electron (i.e., the probability of finding the electron in that region) is different from zero. In other words, if an electron is confined to a room, the quantons will be uniformly distributed in the spacetime world strip of the room as shown in Fig. 6.2. When not measured all quantons will be confined in the room by its walls.

Assume that we have a detector in the room and intend to measure the electron at a given location in the room. We turn on the detector and after some time the electron is registered. So it is *localized* in the detector. The most difficult question is whether *before* the measurement the electron was at the point where the detector is. As quantum mechanics tells us that we can find the electron with *equal* probability in the room, if the electron was at the place of the detector before its registration, then quantum mechanics would be an incomplete theory at best. Let us ask where the electron was one second before its measurement. If the electron were a point-like particle which existed continuously in time, the electron could not be everywhere in the room (where its wave function is different from zero) during that second as seen in Fig. 6.2a. However, if the electron is a 4-atom whose existence in time is discontinuous as shown in Fig. 6.2b, for 1 s it consists of 10^{20} quantons which occupy virtually the whole volume of the room during that time. So, if the electron does not exist continuously in time

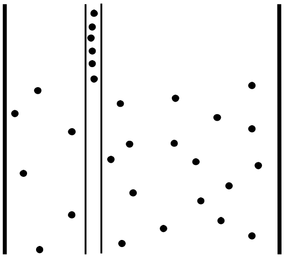


Fig. 6.3. Before being detected, a single electron in a room occupies the whole volume at its disposal. When the first quanton appears in the detector, it is trapped there and all subsequent quantons also start to appear and disappear in the detector

before being measured, it can be everywhere in a volume where the probability of finding it is not zero.

Now let us look at the way a discontinuously existing in time electron is measured. The worldtube of the detector with which we measured the electron in the room is shown in Fig. 6.3. In terms of our three-dimensional language, the quantons of the electron appear and disappear constantly at different points in the room. So before being detected, the electron does occupy the whole volume of the room. If a quanton falls in the detector it is trapped by the walls of the detector (due to a jump in the boundary conditions) and all other quantons will start to appear and disappear only in the detector which means that the electron becomes localized.

The 4-atomistic hypothesis concerning the nature of the quantum object shows a way to resolve the major quantum paradox – a quantum object is always detected as a localized entity, but before detection, it occupies the whole volume where its wave function is not zero. Most importantly, however, this hypothesis suggests that there may be a link between the nature of the quantum object and the nature of spacetime. As shown in Fig. 6.3, the quantons of the electron trapped in a room appear and disappear randomly in the room. In our ordinary three-dimensional language, this can be described only in terms of probability. However, all past, present, and future quantons are *equally* existing in spacetime. We can predict the appearance of a quanton at a given place only probabilistically, but the existence of all quantons is entirely predetermined since they are given at once in spacetime (and Nature 'knows' the locations of all quantons of the electron). So on the 4-atomistic view, quantum probability is objective but not in the sense that the future is open – all quantons comprising an electron are probabilistically distributed in an area of spacetime where the electron wave function is not zero, but they are all equally existent. Therefore the 4-atomistic model of the quantum object not only offers a resolution of the major quantum paradox (and several more as we will see shortly), but also implies that the world is four-dimensional. This is quite a surprising result. The widely held view is that quantum mechanical probabilistic phenomena are incompatible with the four-dimensional static picture of the world deduced from relativity. But if a model of the quantum object which resolves the paradoxes of the wave–particle duality turns out to be in agreement with both the quantum mechanical and the relativistic pictures of the world (and even provides further support for the reality of spacetime), then it deserves careful attention and scrutiny.

It should be emphasized that whether or not the 4-atomistic hypothesis is experimentally confirmed does not really matter for the issue of the nature of spacetime. But it demonstrates that it is not unthinkable to have both – real spacetime and probabilistic behaviour of quantum objects. On the other hand, however, a thorough analysis of the quantum paradoxes reveals that, for their resolution, the nature of the quantum object should be understood and its most probable model appears to be the 4-atomistic model. Here are several reasons for such an expectation.

There are still claims in the literature that quantum mechanics describes ensembles of quantum objects, but cannot be applied to a single electron, for example. The incorrectness of this view is perhaps best demonstrated by the double-slit experiment when *single* electrons are emitted one at a time. The individual electrons confirm the predictions of quantum mechanics but it remains a mystery why the probabilistic quantum laws hold for a *single* electron. The 4-atomistic view provides a nice explanation – the electron itself is an ensemble of entities and 10^{20} quantons for 1 s is quite a good ensemble.

The 4-atomistic view also provides a consistent explanation of the collapse of the wave function. On this view the collapse of the wave function represents a *real* collapse of the quantum object as seen in Fig. 6.3 – after the moment the first quanton appeared in the detector and was trapped there, all subsequent quantons also appear in the detector; that is, the electron collapsed into the detector since its quantons suddenly stopped appearing in the room outside of the detector. But there is no contradiction with special relativity since no superluminal velocities are involved (if the collapsing electron were a fluid-like object, then there would be a contradiction with relativity).

However, a real collapse is not Lorentz invariant. To some this is sufficient to reject any model of the quantum object that involves a real collapse. Such a reaction would not be justified since the electron will be registered by the detector in all reference frames in relative motion. It is true that in some reference frames moving relative to the detector, some of the quantons of the electrons will exist simultaneously with the detection of the electron (the first quanton trapped in the detector), but those quantons cannot be measured, since they belong to the electron which was already detected. So in terms of what can in principle be measured there is no contradiction with the Lorentz invariance requirement. This situation is similar to the one involving virtual particles. Those particles can never be observed directly since they cannot be the end result of any process, but their existence is deduced from their involvement in physical processes. Similarly, some quantons may exist instantaneously with the detection of the electron in some reference frames, but they cannot be detected in any reference frame.

Another source of quantum paradoxes is the so called quantum entanglement – quantum objects appear to be *instantaneously* linked even after they have interacted and separated. On the 4-atomistic view this question is related to the question of transformation of particles. If it turns out that differently charged quantons are the building blocks of all matter, then every quantum object, even those considered indivisible[9] like the electron, consists of a different combination (with a different Compton frequency) of quantons. When a particle transforms into other particles the initial ensemble of quantons splits into sub-ensembles. The quantons of these sub-ensembles may continue to behave as part of the initial ensemble which will be observed as an instantaneous link between the particles produced.

Let me briefly discuss how two other quantum paradoxes – the existence of superpositional states and Schrödinger's cat paradox – look according to the 4-atomistic view. Without having a model of the quantum object, it is impossible to understand how superpositional

[9] To talk about indivisible particles which have constituents appears to be a clear-cut case of contradiction in terms. Hopefully the reader is accustomed to being especially cautious with such 'clear-cut cases' – what appears to be so obviously wrong or right in scientific analyses may not be so in reality. What is promisingly original in the 4-atomistic hypothesis is its radical approach to the way we understand the structure of an object. The present understanding is that an object can have structure only in space. The 4-atomistic model of the quantum object suggests that an object can be indivisible (structureless) in space (like an electron) but *structured in time*.

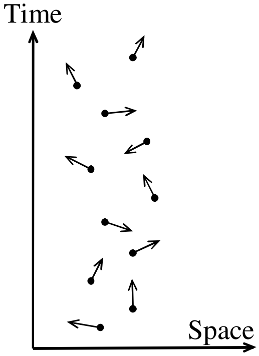


Fig. 6.4. When not observed an electron is represented for one second by 10^{20} quantons, the spins of which are randomly oriented. Therefore during this second the electron really does have all possible spin orientations

states can exist in quantum mechanics. For instant, when an electron is in an external vertical magnetic field and is not measured, it is in a superpositional state of spin up and spin down. It is unclear how the electron can be in both states. The 4-atomistic model does not have any problem with superpositional states – as shown in Fig. 6.4, the spin of every quanton of an electron not in a magnetic field has a different direction. So for any period of time the electron is indeed in a superpositional state of different spins. When the electron is in a vertical magnetic field, half of its quantons have their spins pointing up and half pointing down.

In the Schrödinger cat paradox, a radioactive atom is placed near a detector. When the atom decays, an emitted electron is measured by the detector. The resulting electric current moves a hammer which breaks a glass with poisonous gas. The glass is in a closed box together with a cat. If the atom is intact the glass is not broken and the cat is alive. However, when the atom decays, an emitted electron starts a chain reaction which results in the cat's death. According to quantum mechanics if we have a single atom of half-life one hour, during this hour the quantum object will be in a superpositional state – being *both* intact and decayed. Then the detector, the hammer, the glass, and the cat will all be in superpositional states if we do not open the box during this hour. Opening the box will destroy the superpositional states. The 4-atomistic model of the quantum object also provides a natural solution of the Schrödinger cat paradox. The quantons of the electron that is emitted as a result of the decay of the radioactive atom appear and disappear randomly inside and outside of the atom during the hour. That is why the atom is both decayed and intact during this time. However, no quantons are appearing and disappearing in the detector because if one quanton falls there, the electron will be registered. This means that, on the 4-atomistic hypothesis, the detector cannot be in a superpositional state. Therefore, the hammer, the glass, and the cat cannot be in such a state either. According to the 4-atomistic hypothesis, when the first quanton appears in the detector, it is trapped there, the superpositional state of the decaying atom is destroyed, and all subsequent quantons of the electron start to appear and disappear only in the detector. When the electron is registered, the hammer, the glass, and the cat cannot be in superpositional states, which means that the cat is liberated from being both dead and alive.

6.4 Summary

In this chapter we have addressed the objections against the reality of spacetime according to which special relativity may be regarded as a theory of a four-dimensional world, whilst other physical theories, including some still unconfirmed theories such as quantum gravity and string theory, may tell a different story. We have again stressed that it is the experimental evidence supporting the consequences of special relativity that would be impossible if the world were three-dimensional. Obviously there are two options with such a claim – either it is wrong (and should be proven wrong) or if correct, then the world is really four-dimensional at the macroscopic scale and the interpretations of other theories cannot contradict that result. In this sense special relativity *alone* can provide a solution to the question of the nature of spacetime.

We have also analyzed the apparent contradiction between the static four-dimensional picture of the world depicted by special relativity and the probabilistic nature of quantum mechanics. We have seen that any claim that such a contradiction is genuine should be based on a clear understanding of the nature and origin of probabilities in quantum mechanics; only then would we know whether the probabilistic behaviour of quantum objects contradicts the four-dimensionalist view. As an example to show that it is not unthinkable to have both – real spacetime and objective quantum probabilities – we examined the 4-atomistic model of the quantum object. The conclusion we have drawn is that there might be a link between a model of the quantum object (that resolves the quantum paradoxes) and the dimensionality of the world. So it may turn out that, not only does quantum mechanics not contradict the four-dimensionalist view, but it may provide an independent argument in its support.

7 The Nature of Spacetime and Validity of Scientific Theories

Usually when all arguments against the reality of spacetime fail, the last resort is the reliability of scientific knowledge. This argument comes in two forms. The first is the old philosophical objection against the reliability of all knowledge. Our knowledge relies mostly on inductive reasoning, but since this type of reasoning is not as reliable as deductive reasoning, we cannot be certain about what we know.[1] This philosophical problem is known as Hume's problem of (justification of) induction. The second argument against the reality of spacetime is derived from the widespread view on the validity of scientific theories based on Popper's arguments that a theory can only be disproved; it cannot be proved.

At first sight it appears that one should not bother about such philosophical arguments since it is the *experimental evidence* that supports the four-dimensionalist view. So whether or not inductive inferences can be justified or whether or not special relativity may one day be disproved all seem irrelevant. However, dismissing such arguments without trying to understand the point their authors would like to make would be neither professional nor fair.

A specific instance of the reliability of scientific knowledge argument was discussed in the previous chapter. The idea of this argument is that special relativity does not contain absolute knowledge and for this reason will one day be replaced by another more adequate theory. That theory may provide a different interpretation of the experiments that confirm the relativistic effects and which can now be interpreted

[1] A deductive inference always produces a true conclusion provided that the premises are true. This is so because, in the case of deductive inference, we deduce a statement about an instance of a given phenomenon on the basis of information about the phenomenon as a whole; in other words, in deductive reasoning, the premises entail the conclusion. An inductive inference, however, supports but does not guarantee the truth of the conclusion since this type of inference *extrapolates* knowledge obtained from a given number of examined instances of a phenomenon to the still unexamined instances of the same phenomenon.

only in the framework of the four-dimensionalist view. In this chapter we will address both versions of the reliability of scientific knowledge argument and argue that we have good reason to believe in the unshakable validity of scientific theories in their areas of applicability.

7.1 Reliability of Knowledge: Induction as Hidden Deduction

The first problem we face when we start thinking about the validity of scientific theories is Hume's problem – how to justify inductive inference. As theories are built on postulates which are 'extracted' from experimental observations through the help of inductive reasoning, it is natural to worry whether inductive inferences can be trusted. What gives us confidence in induction is the fact that all scientific theories obey the requirements of the hypothetico-deductive method:

- using experimental facts and inductive reasoning to formulate postulates (hypotheses),
- deducing predictions from them,
- returning to experiment to test the predictions derived from the postulates.

Every time experiment confirms a given prediction, our faith in the postulates increases, but it is not impossible to discover an experimental fact that contradicts them. If that happens we have no choice but to re-examine and eventually replace the contradicted postulate (or postulates). As the postulates are *inductive* inferences, it is exactly in cases like these that the issue of justification of induction arises.[2]

Since Hume there have been many attempts to justify induction, but it seems one possibility has never been explored. In this section I will briefly argue that, ultimately, induction turns out to be *hidden deduction*, which explains why all statements based on correct inductive reasoning have never been found wrong.

First, let me explain what I mean by a correct inductive inference. At sea level water boils at 100°C and it seems natural to extrapolate this piece of knowledge inductively to all unexamined instances of boiling water by saying that it boils at 100°C. But on Mount Everest, for instance, water boils at a lower temperature and one may be tempted

[2] This brief description of the hypothetico-deductive method represents, of course, an idealized situation. But it is sufficient to demonstrate the problem with inductive inferences.

to say that this is an excellent example of how induction fails. However, it is clear that such an inductive inference is incorrect, since the conditions (e.g., the atmospheric pressure) at sea level and on Mount Everest are not the same. Similarly, when we say all observed ravens have been black and therefore all ravens are black, we again fail to apply inductive reasoning correctly. As in the first example the conditions involved in the observation of a black and, say, a white raven are not the same – the expressed genetic material is not the same for a black and white raven. To my knowledge no case of a correct inductive inference that fails has ever been observed.

Hardly anyone questions the reliability of our knowledge, which is essentially obtained through inductive reasoning.[3] The fundamental question is why inductive reasoning produces true results. To outline an answer to this question consider the inductive inference: "since the humans who lived so far died, all humans are mortal", the truth of which no one doubts. Our knowledge of why humans are mortal reveals that being mortal is one of the intrinsic features that define a human; therefore mortality is *contained* in the definition of a human. That is why, when we say 'all humans are mortal', we are employing *deductive* reasoning, since the definition of a human entails the conclusion that every being that fits that definition is mortal.

Once we have comprehensive knowledge of the basic characteristics of an object or a phenomenon obtained through the study of a *limited* number of objects or phenomena from the same class, we are not applying inductive inference when claiming that all objects or phenomena from that class possess those basic characteristics. It is a *deductive* inference that all objects from a given class possess the features which are contained in the definition of the objects from that class. For example, electrical conductivity is an intrinsic characteristic of metals and therefore is contained in the definition of a metal.[4] For this reason the

[3] Even representatives of the extreme philosophical skepticism trust inductive reasoning – for instance, they take their umbrella if it is raining outside.

[4] Here is a typical definition of a metal from a physics encyclopedia [54]: "The properties that essentially characterize a metal are its high electrical and thermal conductivity, its ductility and malleability, and its luster. These properties may in part be deduced from the type of binding that characterizes a metal." As far as electrical conductivity is concerned, it is fully deduced from the atomic structure of metals, which shows that electrical conductivity is indeed a defining feature of all metals. What is crucial is to realize that our knowledge of the atomic structure is not obtained through induction. Obviously, we do not examine the atomic structure of a hundred atoms and extrapolate the knowledge gained to all atoms.

conclusion ‘all metals conduct electricity’ is obviously derived through deductive reasoning, since every metal, by definition, conducts electricity. On the other hand, we cannot say all ravens are black since we know from genetics that blackness is not an intrinsic feature of ravens.

When we gain sufficient knowledge about the intrinsic features of a class of objects or phenomena, it becomes clear why induction works and why it appears to fail. If a given instance of inductive reasoning fails, it is found upon a closer examination that it has been incorrectly applied to cases which are not equal in all respects to the examined cases on the bases of which the inductive inference has been made. In all cases in which an apparently inductive inference has been confirmed experimentally (e.g., measuring the electrical conductivity of an unexamined piece of metal), it turns out to be hidden deductive reasoning.[5] So what appears to be inductive reasoning is reliable since it is in fact deductive reasoning, which cannot lead to wrong conclusions provided that the premises are true.

It is true that the postulates of a theory are formulated on the basis of inductive reasoning. An example are Newton’s three principles. As we will see in the next section, they will forever remain correct in the area where their predictions have been experimentally confirmed. So again, the question is why do correct inductive inferences always produce reliable results? The answer is the same – all correct inductive inferences turn out to be instances of hidden deduction. In the case of Newton’s three principles (and the postulates of all accepted scientific theories), it happened that intrinsic features of physical objects had been correctly *guessed*. Such a guess, however, is not induction. Inductive reasoning merely extrapolates what has been observed in all examined instances of a phenomenon to all unexamined ones. No attempt is made to understand what the intrinsic (defining) features of that phenomenon are. On the contrary, any postulate of a scientific theory attempts to capture the most fundamental features of the phenomena that the theory describes. At a later time, when the phenomena have been thoroughly studied, it does turn out that the

[5] Some may object to this conclusion and may argue that we arrive at the basic characteristics of an object (say, a metal) through induction. This is not the case since we learn about these characteristics, not by examining different samples of that object, but by studying the *structure* of the object. Similarly, when we say that mortality is contained in the definition of a human it may be argued that we arrived at that conclusion by induction. This argument could have been forcefully advanced in the 19th century. But now the inclusion of mortality in the definition of a human is based not on induction, but on the latest achievements of genetics.

postulates of the theory of those phenomena are merely stating their basic properties. That is why the application of the postulates of the theory to still unexamined instances of the same phenomena is always successful. Such an application is simply a deductive inference – as the unexamined instances are manifestations of the same phenomena, they should look exactly like the already examined ones. Newton succeeded in reflecting all intrinsic properties of physical bodies (excluding gravity) in his three principles. This explains why all unexamined bodies behave according to his principles, no matter where in the universe they are. The application of Newton's principles is really deduction – as inertia is one of the defining features of a physical body, all still unexamined physical bodies must possess inertia, which means that:

- they should move on their own if nothing prevents them from doing so (Newton's first principle),
- they should resist when something tries to deviate them from moving by inertia (Newton's second principle),
- the resistant (inertial) force is equal to the external force that *caused* the resistant force (Newton's third principle).

Now special relativity appears to have provided us with sufficient knowledge on the basis of which we can be certain that Newton's three principles are really stating the fundamental features of *all* physical bodies when gravity is not taken into account. A physical body is a worldtube in spacetime. If there is just one worldtube in a given spacetime region, it is straight in Minkowski spacetime or geodesic in a curved spacetime. In three-dimensional language this means that the straight worldtube is perceived by us as a three-dimensional body which moves on its own with constant velocity (in the case of Minkowski spacetime). As a free body's motion by inertia is a manifestation of the fact that the body's worldtube is straight, we now know why Newton's first principle works – having captured the concept of inertia, it reflected the fact that the worldtube of a free body is a straight 'line' in Minkowski spacetime. When the worldtube of a body is curved by another worldtube it resists the deformation and, as we will see in Chap. 10, the resistant (inertial) force has the form of Newton's second law. When two worldtubes mutually deform each other, each of them resists its deformation caused by the other worldtube, which shows that what is an external force with respect to one worldtube is a resistance force for the other worldtube and vice versa. Due to this symmetry the two forces have equal magnitudes and opposite directions, which is Newton's third principle. So, according to special relativity, Newton's three principles are simply statements that there are straight worldtubes in Minkowski spacetime which, like ordinary tubes or rods, resist when deformed.

I believe it is clear that the way the problem of justification of induction was formulated and discussed here excludes many cases that are usually regarded as instances of inductive reasoning. For example, if a hundred transatlantic flights have been safe it is believed that we apply inductive reasoning when we expect the hundred and first flight to be safe as well. But if it is not, then should we assume that inductive inferences are unreliable? It is evident that the hundred and first flight was *different* from the other flights as far as the factors responsible for the safety of a flight are concerned. In situations like this there are a lot of factors involved and it is virtually impossible to regard all instances as equal in all relevant respects; in this sense the first hundred flights are not equal either. For this reason in similar cases, one has two options – either to redefine inductive inference in order that it involves not equal examined and unexamined cases, or introduce a *probabilistic inference* to cover all cases where the examined and unexamined instances are not equal in all respects.[6] A careful examination of the way induction works in science (especially in physics) demonstrates that it is crucial for any inductive inference to involve only examined and unexamined instances which are equal in all relevant aspects. If this were not the case, we would not be able to talk about the same law of nature, say, in the case of boiling water. Therefore one should reserve the concept of inductive inference for examined and unexamined instances of the *same* phenomenon or the *same* class of objects. In all other cases, probabilistic inferences should be employed.

As science, including special relativity, deals with inductive inferences, we should obviously be concerned about the reliability of this type of inference. Our analysis has shown that we can trust it, since correct inductive inferences are, in fact, hidden deductive inferences and are therefore as reliable as deductive inferences themselves. So the first form of the reliability of scientific knowledge argument has been addressed and we have seen that even this philosophical argument

[6] Another example of such a case is the following. If the first hundred examined citizens of a small German town are German it appears *probable* to conclude that all citizens of the town are German. If the hundred and first citizen is not German, it is clear that this instance is not equal in all relevant respects to the first hundred instances; that person might have moved to the town from another country.

does not challenge our trust in special relativity. This means that its implications for the reality of spacetime hold.

The second form of the reliability of scientific knowledge argument applies to cases of experiments that contradict a scientific theory and its postulates. When a postulate is contradicted by an experiment, should we conclude that it was a result of incorrectly applied inductive reasoning? An incorrect inductive inference may be blamed for the contradiction in cases when a new theory is being developed and different postulates are tested. However, this is unlikely if such a contradiction is discovered in an *accepted* theory, which means that the postulate has been repeatedly tested. Then it appears that the only option that remains to explain the contradiction is a failure of induction. So is the analysis carried out in this section flawed? At this point it will be rewarding to assume that this is really the case and try to find the flaw. We will see in the next section that, in the case of a theory whose predictions have been repeatedly confirmed by experiment, a contradiction of a given prediction with an experimental fact does not necessarily imply that a postulate (and therefore the inductive inference used in its formulation) is wrong in the area where the predictions have been successfully tested.

7.2 Correspondence Principle and Growth of Scientific Knowledge

Popper believed that if a prediction of a theory is contradicted by an experiment it necessarily means that a postulate is contradicted and therefore the theory is wrong. This is quite obvious and perhaps for this reason the widespread view on confirmation and validity of scientific theories holds that any theory may be proven wrong. As Popper put it, the highest status a scientific theory can attain is not yet disconfirmed. If this view were correct then those who are not worried by the implications of special relativity might have a good point – one day a new theory might replace special relativity and a new interpretation of the experimental evidence confirming the relativistic predictions might be found.

However, if such a view implies that an accepted theory, one of whose predictions is later found to contradict experiment, is wrong and should be abandoned, then it does not reflect the real situation in science. For instance, predictions derived from the equations of motion of special relativity clearly contradict the existing quantum mechanical experimental evidence. But no one declares special relativity a wrong theory. When a prediction of a theory, whose other predictions have been repeatedly and successfully tested, is found to contradict some new experiments, the most likely explanation is that the theory has been applied outside of its area of applicability. Such an explanation does not question the validity of the theory in the area where its predictions have been experimentally confirmed. In the example given above the failure of the relativistic equation of motion to describe the motion of quantum objects is a result of applying special relativity outside of its area of applicability.

As the issues of confirmation and validity of scientific theories are of direct importance for the debate over the nature of spacetime, let as examine briefly in what sense an accepted theory can be wrong. We will do this by making use of the correspondence principle. It was first formulated for the case of quantum phenomena and applied to the theory of atomic structure by Bohr in 1923. In the transition from microscopic to macroscopic levels, where classical and quantum theories should agree, quantum mechanics must reduce to the classical theory. In the case of the Bohr model of the atom, classical and quantum mechanics agree when the energy difference between quantized allowed energy levels is very small for very large quantum numbers n.

The correspondence principle turned out to be a universal principle – any new theory (whatever its character) should reduce to the previous, well-established theory to which the new theory corresponds when applied to the area where the less general theory is known to hold. Put in another way, any new theory should contain the previous one as a limiting case. Here is a typical scientist's statement about the fate of old scientific theories [55]: "Quantum mechanics [...] doesn't displace Newtonian mechanics, but incorporates it as a limit. Scientific theories grow by incorporating what is already known and adding to it [...]." This is correct, but the correspondence between different theories of modern physics is more complex. Here are two examples of correspondence between theories which reveal different aspects of the correspondence principle:

- The correspondence between special relativity and Newtonian mechanics when the speeds of objects are slow (compared to the speed of light) – special relativity coincides with Newtonian mechanics in the limit $v/c \rightarrow 0$ (with the exception of the expression for the rest energy $E = mc^2$ of a particle, which does not have an analog in pre-relativistic physics).
- The correspondence between quantum mechanics and Newtonian mechanics when quantum mechanics is applied to macroscopic re-

> gions where Newtonian mechanics operates. In other words, quantum mechanics coincides with Newtonian mechanics when a physical quantity of the objects under study called the action (mass × speed × distance) is large compared to Planck's constant $\hbar$, which is true in the macroscopic world but not at the atomic scale.

These two examples illustrate two different aspects of the correspondence principle. As special relativity and Newtonian mechanics describe the *same* (macroscopic) level of reality, the correspondence between them does not have any ontological counterpart. It simply reflects the deepening of our knowledge about the macroscopic world – special relativity is a better (more precise) theory than Newtonian mechanics – and, therefore, the correspondence between special relativity and Newtonian mechanics has mostly epistemological content. That is why the correspondence between special relativity and Newtonian mechanics demonstrates only one side of the correspondence principle – its epistemological aspect.

However, the situation is quite different when the correspondence between quantum mechanics and Newtonian mechanics is considered. In this case the ontological aspect of the correspondence principle is at work, since the correspondence between quantum mechanics and Newtonian mechanics has a clear ontological content – it reflects the correspondence between quantum and macroscopic physical laws. However, since we express that correspondence in terms of our knowledge (our theories – quantum mechanics and Newtonian mechanics), the correspondence between quantum and macroscopic laws inevitably contains epistemological elements as well.

The epistemological aspect of the correspondence principle operates at a *single* level of the structural organization of matter – it requires that any new theory of the *same* level contains the preceding theories as limiting cases. This means that the new theory should reduce to the already existing theory in the area where the old theory is working properly.

The ontological aspect of the correspondence principle applies to theories describing neighboring levels, such as quantum mechanics and Newtonian mechanics. When a new theory describing a more fundamental level (lying 'below' the level at which the existing theory operates) is formulated, the correspondence principle requires that, when applied to the 'upper' level, the new theory should reduce to the theory of that level.

The present status of our knowledge appears to indicate that the ontological aspect of the correspondence principle works in a one-way fashion – it requires that the more fundamental theory should reduce to the theory of the 'upper' level, but not vice versa. Another feature of the ontological aspect that needs clarification concerns the correspondence between theories describing neighboring levels, each of which is described by more than one theory. For example, should quantum mechanics as the first theory of the quantum level reduce to the first macroscopic theory (Newtonian mechanics) or to the second, more precise macroscopic theory (relativity)? Or, in general, should the latest theory of a given level reduce to the latest theory describing the 'upper' level?

There are other open questions involving both the epistemological and ontological aspects of the correspondence principle which need more research. For instance, the incommensurability (translatability) issue should be addressed in both aspects of the correspondence principle. The epistemological aspect of the correspondence between the Newtonian gravitational theory and general relativity deserves special attention. One can find a correspondence between the equations of the two theories, but it is the correspondence between the systems of concepts in the theories that needs to be clarified. For example, how the correspondence between *gravity as a force* (in the Newtonian case) and *gravity as spacetime curvature* (in general relativity) should be understood. The need to address the incommensurability thesis in the case of the ontological aspect of the correspondence principle is even more urgent – how should we understand the correspondence between the strange quantum world and our ordinary world?

The analysis of the correspondence principle shows that scientific knowledge grows in two major ways:

- by a more precise description of the same level of the world,
- by describing new levels of reality.

Distinguishing between the two aspects of the correspondence principle makes it possible to address at least two important questions, the first of which is of direct relevance for the nature of spacetime:

- Can an accepted scientific theory be refuted?
- Is a final scientific theory possible?

7.3 Can an Accepted Scientific Theory Be Refuted?

The epistemological aspect of the correspondence principle clearly indicates that, whenever an accepted theory is replaced by a new one, the old theory remains correct in its area of applicability, where its predictions have been *experimentally* confirmed. When applied to phenomena lying outside of that area, the old theory naturally fails. So, a scientific theory is not wrong in an absolute sense as is sometimes claimed – it gives wrong predictions only when forced to work outside of its area of applicability. This shows that, whenever a postulate of an accepted theory is contradicted by an experiment, it is still correct where the theory works, which means that the inductive inferences used in its formulation are as reliable as the deductive inference used to extract predictions from the postulate.

Thus the correspondence principle demonstrates that a scientific theory cannot be proven wrong in its area of applicability (where its predictions have been experimentally confirmed). If you think this is too strong a claim, imagine how many people might doubt that, a thousand years from now, classical mechanics will still be applied when bridges are built (if we still need them then) and classical electrodynamical calculations will be used for the wiring of buildings (if electricity is still used then). That is why it is unrealistic to expect special relativity to be disproved one day in its area of applicability, where its predictions have been experimentally confirmed.

As it is the new theory that defines the area in which the previous theory is correct, it appears possible (but quite challenging) to try to define the area of applicability of an existing theory *before* the arrival of a new one. A good starting point might be to try Dostoevsky's method of cruel experimentation [56]. By putting his characters in extreme situations, he lets them reveal their true nature. In his special relativity, Einstein applied the same method – Nature revealed an aspect of her true nature when the extreme case of objects moving at high speeds was considered. By letting the speeds of objects take on greater and greater values, it is now clear that classical mechanics leaves its area of applicability, where its predictions were experimentally confirmed, since its predictions are no longer accurate for extremely high speeds. As infinities by their very nature constitute extremes, it is evident that an extreme situation arises whenever there are infinities in a theory. That is why an indication that a theory may be claiming to cover phenomena lying beyond its area of applicability might be the allowance of infinities by the theory. For example, Newtonian mechanics allows motions with infinite velocities and it is precisely there that it fails. Newtonian physics also allows any physical quantities to take on continuously smaller and smaller values (infinite divisibility), and this is another area where it fails.

7.4 Is a Final Scientific Theory Possible?

The epistemological aspect of the correspondence principle appears to imply that it is natural to expect that one day we may formulate a final theory which describes a given level of the world. The reason is that any level is described in terms of a *finite* number of kinds of objects and physical laws, and therefore it cannot be expected that a theoretically infinite number of theories will be necessary to describe adequately the objects and laws of a *single* level of the world. Why a final theory describing a given level of reality appears to be logically unavoidable is evident from the following argument. As any new theory reduces the area of applicability of the previous theories of the same level, an infinite number of theories will reduce the area of applicability of the first theory to zero, which means that the first theory will turn out to be an absolutely wrong theory. That would make the experimental confirmation of its consequences not merely a mystery, but perhaps a crisis in science.

The prospect of a final theory, however, does not imply an end to science. The ontological aspect of the correspondence principle contains the possibility of an infinite growth of our knowledge. If there exists an infinite number of levels of the world, we will obviously need an infinite number of theories to describe everything that exists.

7.5 Summary

This chapter has been concerned with the two forms of the reliability of scientific knowledge argument against the validity of special relativity. We have seen that the argument fails. Inductive inferences can be trusted as much as deductive inferences, since they are hidden deductive inferences. Therefore the validity of special relativity cannot be questioned on the basis that inductive inferences are unreliable.

The second form of the philosophical argument against the validity of special relativity holds that any theory may in principle be disproved. An examination of the validity of a scientific theory in terms of the correspondence principle has revealed several important results:

- a scientific theory cannot be disproved in its area of applicability, where its predictions have been experimentally confirmed (and hence special relativity will remain correct in the area where it has been successfully experimentally tested);

- there are two major ways in which scientific knowledge grows – by a more precise description of the same level of the world and by describing new levels of reality;
- it is not unrealistic to expect that every level of the world can be described by a final scientific theory which, however, does not imply an end to science, since the remaining levels of the world will be described by different theories.

8 Propagation of Light in Non-Inertial Reference Frames

8.1 Acceleration Is Absolute in Special and General Relativity

We have seen in Chap. 4 that special relativity, which describes the physics of flat spacetime, provides a clear criterion for the absoluteness of acceleration in flat spacetime – the worldline of an object moving with constant velocity is a *straight* line, whereas the worldline of a body subjected to the ordinary (flat-spacetime) acceleration

$$a^{\mu}_{\text{flat}} = \frac{\mathrm{d}^2 x^{\mu}}{\mathrm{d}\tau^2}$$

is curved. However, there are no straight worldlines in curved spacetime, which is described by general relativity. As a straight worldline in flat spacetime represents a body moving non-resistantly (i.e., by inertia) the same requirement is used in curved spacetime to define a special class of worldlines representing bodies whose motion is non-resistant (i.e., by inertia) – such worldlines are called *geodesics*. In other words, the worldline of a body, whose curved-spacetime acceleration

$$a^{\mu}_{\text{curved}} = \frac{\mathrm{d}^2 x^{\mu}}{\mathrm{d}\tau^2} + \Gamma^{\mu}_{\alpha\beta}\frac{\mathrm{d}x^{\alpha}}{\mathrm{d}\tau}\frac{\mathrm{d}x^{\beta}}{\mathrm{d}\tau}$$

is zero, is a geodesic. Then the same criterion for the absoluteness of acceleration holds in curved spacetime as well: a body whose worldline is geodesic is not subjected to a curved-spacetime acceleration (i.e., $a^{\mu}_{\text{curved}} = 0$) and moves non-resistantly (by inertia), whereas the worldline of a body whose curved-spacetime acceleration a^{μ}_{curved} is different from zero is not geodesic.[1]

[1] In contrast to the situation in special relativity, where acceleration is absolute, in general relativity there exists a *relative acceleration* in addition to the absolute acceleration defined here. As there are no straight worldlines in curved spacetime, two bodies whose worldlines are geodesic will appear to accelerate relative to each other; the rate of change of the distance between them is given by the equation of geodesic deviation [57, p. 343].

Therefore, the absolute difference between a geodesic worldline (which in the case of flat spacetime represents a body whose absolute uniform motion cannot be detected) and a non-geodesic worldline makes accelerated motion absolute. Such a conclusion appears to imply that there is *one* space. Why this is not so was discussed in Chap. 4. As all observers in relative motion have different spaces, an accelerating body moves relative to those spaces, not relative to a *single* space. Acceleration is absolute in the sense that a curved worldline is curved for all observers, which explains why the accelerated motion of a body can be discovered from within the non-inertial reference frame in which the body is at rest. It follows from the analysis of the nature of spacetime carried out in Chap. 5 that this difference is not just a feature of the mathematical formalism of special and general relativity which may have no bearing on how acceleration should be understood. The worldlines of the physical bodies are real four-dimensional objects, according to special relativity, and more importantly according to the experimental evidence that confirms its predictions. This means that the difference between a geodesic and a non-geodesic worldline reflects an objective fact in the external world which can be discovered experimentally. In other words, we should be able to detect the motion of a non-inertial reference frame, in which an accelerating body is at rest, by carrying out experiments within that frame.[2] This implies that the forms of the laws of nature in inertial and non-inertial reference frames are not the same. An immediate consequence from here is that the ve-

[2] In flat spacetime, a non-inertial reference frame is associated with an accelerating body, say a spacecraft whose engines are working. In curved spacetime, however, a non-inertial reference frame is associated either with an accelerating body (like such a spacecraft), whose worldline is not geodesic, or with a body at rest in a gravitational field, whose worldline is not geodesic either. In both cases the body's curved-spacetime acceleration $a^{\mu}_{\text{curved}} \neq 0$. In flat spacetime an inertial reference frame is associated with a body moving by inertia. The situation in curved spacetime, however, is more complicated. Perhaps the best way to think of whether an inertial reference frame can be introduced is by asking whether or not Cartesian coordinates can be used. A Cartesian coordinate system (whose axes are straight lines) can be introduced *globally* in flat spacetime and represents a global inertial frame; this can be done since there are straight lines in flat spacetime. Due to the non-existence of straight lines in curved spacetime, Cartesian coordinates cannot be introduced there, which means that one cannot talk about global inertial frames in curved spacetime. However, Cartesian coordinates can be used in the infinitesimal neighborhood of any given point in curved spacetime, since the infinitesimal neighborhood at the given point can be regarded as a small flat spacetime region that is tangent to the curved spacetime at that point; in that region a geodesic worldline is a straight line. For this reason one can introduce only local inertial frames in curved spacetime.

locity of light is not constant in non-inertial frames – it depends on the frame's proper acceleration.[3] This dependence allows a non-inertial observer to detect his accelerated motion by using light signals.

8.2 The Need for Two Average Velocities of Light in Non-Inertial Reference Frames

So far, the corollary of relativity that the velocity of light determined in a non-inertial reference frame depends on the frame's acceleration has received little attention. As a result it has not been realized that two average velocities of light – an average coordinate velocity and an average proper velocity – are needed for a complete description of the propagation of light in non-inertial reference frames. Both average velocities depend on the proper acceleration of the non-inertial reference frame. It should be stressed, however, that it is the average coordinate velocity of light *between two points* that is different from c; the local speed of light measured at a point is always c. Let us consider several examples that demonstrate why the average light velocities are needed.

- Einstein's thought experiment [58] involving an elevator at rest in a *parallel* gravitational field[4] of strength g and an elevator which accelerates with an acceleration $\boldsymbol{a} = -\boldsymbol{g}$ was designed to demonstrate the equivalence of the non-inertial reference frames N^{g} (associated with an elevator at rest in the gravitational field) and N^{a} (associated with an elevator accelerating in a space devoid of gravity). Einstein called this equivalence the principle of equivalence: it is not possible by experiment to distinguish between the non-inertial frames N^{a} and N^{g}, which means that all physical phenomena look the same in N^{a} and N^{g}. Therefore if a horizontal light ray propagating in N^{a} bends, a horizontal light ray propagating in N^{g} should bend as well. What is most important for the question of the absoluteness of acceleration in Einstein's thought experiment is that the bending of light in an accelerating reference frame N^{a} is precisely an effect that allows an observer in N^{a} to conclude that he is accelerating. However, the bending of light does not immediately demonstrate that an average velocity of light should be used to

[3] The acceleration of a non-inertial reference frame determined in the instantaneous (comoving) inertial frame is called proper acceleration.

[4] For the issues discussed here it is sufficient to assume that the gravitational field in an elevator at rest on the Earth's surface is a good approximation of a parallel gravitational field.

describe the propagation of light in N^{a}. To see that such a velocity is indeed needed, assume that, instead of a horizontal ray, the observer in N^{a} decides to use two vertical rays[5] – one emitted from the elevator ceiling downwards and the other from the elevator floor upwards (as shown in Fig. 8.1). When the accelerator is moving with constant velocity, the two light rays will meet at the middle point B (between the floor and the ceiling). However, when the elevator accelerates the rays will meet at a point B' situated below B in the direction of the elevator floor. As we will see in the next section, this fact cannot be explained without the introduction of an average *coordinate* velocity of light which depends on the elevator's proper acceleration.

- An observer in a rotating reference frame, a rotating disk for example, can also detect the disk's accelerated motion by using light (the so-called Sagnac effect): light signals emitted from a point P in opposite directions along the rim of the disk do not arrive at the same time at P [63]. Without taking into account the fact that the velocity of light as determined in the rotating reference frame depends on the frame's acceleration, this effect cannot be adequately explained either.
- A second average velocity of light – an *average proper* velocity of light – is required to explain a number of phenomena in which the velocity of light is determined with respect to a *given* point. For instance, the average proper velocity of light is implicitly used in the Shapiro time delay [64, 65]. It also turns out that this is not always c. The fact that it takes more time for a light signal to travel between two points P and Q in a gravitational field than between the same points in flat spacetime as determined by an observer at one of the points indicates that the average velocity of light between the two points is smaller than c. Unlike the average coordinate velocity, the average proper velocity of light between two points depends on which point it is measured at. This fact confirms the dependence of the Shapiro time delay on the point where it is measured and shows, as we shall see in Sect. 8.5, that in the case of a parallel gravitational field, it is not always a delay effect (in such a field the average proper velocity of light is defined in terms of both

[5] Although even introductory physics textbooks [59–62] have started to discuss Einstein's elevator experiment, the following obvious question has been overlooked: Are light rays propagating in an elevator in a vertical direction (parallel and anti-parallel to $\boldsymbol{a}$ or $\boldsymbol{g}$) also affected by the accelerated motion of the elevator or its being in a gravitational field?

the proper distance and proper time of an observer, which justifies use of the term ‘proper’). A light signal will be delayed *only* if it is measured at a point P that is farther from the gravitating mass producing the parallel field; if it is measured at the other point Q closer to the mass, it will take less time for the signal to travel the same distance. This shows that the average proper velocity of the signal determined at Q is greater than that measured at P and slightly greater than c. As we will see in Chap. 9, the average proper velocity of light is also needed if we are to calculate the potential and electric field of a charge in a non-inertial reference frame *directly* in that frame, without the need to transform the field from the local inertial reference frame.

- The introduction of the average velocities of light also sheds some light on a subtle feature of the propagation of light in the vicinity of a massive body – whether or not light falls in its gravitational field. The particle aspect of light seems to entail that a photon, like any other particle, should fall in a gravitational field (due to the mass corresponding to its energy) and the deflection of light by a massive body appears to support such a view. And indeed this view is sometimes implicitly or explicitly expressed in papers and books, although the correct explanation is given in books on general relativity (see for instance [22,66]). It has been claimed recently that the issue of whether or not a charge falling in a gravitational field radiates can be resolved by assuming that the charge’s electromagnetic field is also falling [67]. Such a claim needs a detailed justification, since an electromagnetic field falling in a gravitational field implies that light falls in the gravitational field as well, which is not the case as we will see shortly. Even Einstein and Infeld appear to suggest that, as a light beam has mass on account of its energy, it will fall in a gravitational field [58]: “A beam of light will bend in a gravitational field exactly as a body would if thrown horizontally with a velocity equal to that of light.” This comparison is not quite accurate, since the vertical component of the velocity of the body will increase as it falls, whereas the velocity of the ‘falling’ light beam is decreasing for a non-inertial observer (supported in a gravitational field), as we shall see below. Sometimes statements such as “a beam of light will accelerate in a gravitational field, just like objects that have mass” and therefore “near the surface of the earth, light will fall with an acceleration of 9.81 m/s^2” can be found in introductory physics textbooks [59]. We shall see later that during its 'fall' in a gravitational field, light is slowing down – a negative acceleration of 9.81 $\mathrm{m/s^2}$ is decreasing its velocity.

Before deriving the average velocities of light in non-inertial reference frames, I would like to comment on a possible explanation of why the need to introduce these velocities has been overlooked. The coordinate velocity of light was first used in general relativity, where the physical significance of coordinate-dependent quantities is often regarded as doubtful. What perhaps reinforces the reluctance to take the coordinate velocity of light more seriously is the fact that it is a function of the gravitational potential. As the gravitational potential is determined to within a constant, it therefore appears that the coordinate velocity of light does not reflect a physical quantity.

Leaving aside the fact that the same argument applies to the gravitational potential itself (and I doubt that many physicists will put their names under a statement that the gravitational potential does not reflect anything objective), I will outline several arguments which, in my view, show that the coordinate velocity of light deserves closer attention in general relativity and in non-inertial reference frames in general:

- The coordinate velocity of light was used by Einstein in his 1916 paper to calculate the deflection angle of light bending near the Sun [69] (see also [70]).
- The coordinate velocity of light is also used to calculate the retardation of light in a gravitational field (Shapiro time delay effect) [57, p. 197], [66, p. E-1].
- The very concept of black holes implies that the coordinate velocity of light reflects an important fact about these objects – the coordinate velocity of light at the event horizon is zero which explains why light cannot escape from the region of extreme spacetime curvature surrounding a black hole.
- The fact that light signals propagating along and against the acceleration of a non-inertial reference frame (discussed in the next two sections) meet at B' (not B) cannot be explained without the introduction of an average coordinate velocity of light which is not equal to c.
- In an accelerating reference frame, the coordinate velocity of light depends on the frame's proper acceleration, which is not determined to within a constant. In a parallel gravitational field the coordinate velocity of light also depends on the frame's proper acceleration.

- The coordinate velocity of light is used in the derivation of the average proper velocity of light, which in turn is needed for the calculation of (i) the Shapiro time delay and (ii) the potential, electric field, and the self-force of a charge directly in a non-inertial reference frame (as we will see in Chaps. 9 and 10).
- The most radical and the most convincing argument for the use of the coordinate velocity of light, in my view, comes from the issue of the nature of spacetime. It is true that the definition of gravitational potential contains an element of convention which therefore makes the definition of the coordinate velocity of light conventional to some extent as well. Obviously, what is important here is the consistent use of a convention once it is accepted; that is why the concept of gravitational potential works in physics. We have seen in Chap. 5 that the true reality is a four-dimensional world in which there is no such thing as velocity – there are only worldlines. Only when we want to describe this timelessly existing world in terms of our everyday three-dimensional language can we introduce the concept of velocity. But since this concept does not have an ontological counterpart and only serves the purpose of *description*, it does follow that we are free to choose the language in which we describe the external (velocity-free) world.

8.3 Average Coordinate Velocity of Light

The calculation of the average coordinate velocity of light between two points in an accelerating reference frame N^{a} can be carried out by considering two extra light rays parallel and anti-parallel to the acceleration $\boldsymbol{a}$ of the Einstein elevator, in addition to the horizontal ray originally considered by Einstein.

Consider a non-inertial reference frame N^{a} in which an elevator accelerating with acceleration $a = |\boldsymbol{a}|$ is at rest (Fig. 8.1). Three light rays are emitted *simultaneously* in the elevator (in N^{a}) from points D, A, and C toward point B. Let I be an inertial reference frame instantaneously at rest with respect to N^{a} (i.e., the instantaneously comoving frame) at the moment the light rays are emitted. As I and N^{a} have a common instantaneous three-dimensional space and therefore common simultaneity at the moment the three light signals are emitted, the emission of the rays is simultaneous in N^{a} as well as in I. At the next moment an observer in I sees that the three light rays arrive simultaneously not at point B, but at B', since for the time $t = r/c$ the light rays travel toward B, the elevator moves a distance

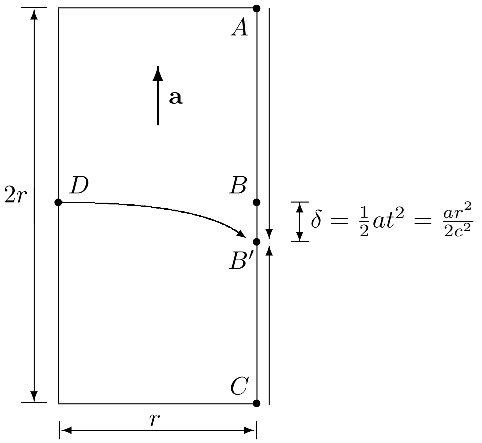


Fig. 8.1. Three light rays propagate in an accelerating elevator. Having been emitted simultaneously from points A, C, and D, the rays meet at B'. The ray propagating from D toward B, but arriving at B', represents the original thought experiment considered by Einstein. The light rays emitted from A and C are introduced in order to determine the expressions for the average velocity of light in an accelerating frame of reference

$\delta = at^2/2 = ar^2/2c^2$. As the simultaneous arrival of the three rays at point B' as viewed in I is an absolute (observer-independent) fact due to its being a *point* event, it follows that the rays arrive simultaneously at B' as seen from N^{a} as well. Since for the *same* coordinate time $t = r/c$ in N^{a}, the three light rays travel *different* distances $DB' \approx r$, $AB' = r + \delta$, and $CB' = r - \delta$, before arriving simultaneously at point B', an observer in the elevator concludes that the propagation of light is affected by the elevator's acceleration. The *average* velocity $c^{\mathrm{a}}_{AB'}$ of the light ray propagating from A to B' is slightly greater than c:

$$c^{\mathrm{a}}_{AB'} = \frac{r+\delta}{t} \approx c\left(1 + \frac{ar}{2c^2}\right) .$$

The average velocity $c^{\mathrm{a}}_{B'C}$ of the light ray propagating from C to B' is slightly smaller than c:

$$c^{\mathrm{a}}_{CB'} = \frac{r-\delta}{t} \approx c\left(1 - \frac{ar}{2c^2}\right) .$$

It is easily seen that to within terms proportional to c^{-2} the average light velocity between A and B is equal to that between A and B', i.e., $c^{\mathrm{a}}_{AB} = c^{\mathrm{a}}_{AB'}$ and also $c^{\mathrm{a}}_{CB} = c^{\mathrm{a}}_{CB'}$:

$$c^{\mathrm{a}}_{AB} = \frac{r}{t - \delta/c} = \frac{r}{t - at^2/2c} = \frac{c}{1 - ar/2c^2} \approx c\left(1 + \frac{ar}{2c^2}\right) \tag{8.1}$$

and

$$c^{\mathrm{a}}_{CB} = \frac{r}{t + \delta/c} \approx c\left(1 - \frac{ar}{2c^2}\right) . \tag{8.2}$$

As the average velocities (8.1) and (8.2) are not determined with respect to a specific point and since the *coordinate* time t is involved in their calculation, it is clear that the expressions (8.1) and (8.2) represent the average *coordinate* velocities between the points A and B and the points C and B, respectively.

The same expressions for the average coordinate velocities c^{a}_{AB} and c^{a}_{CB} can also be obtained from the expression for the coordinate velocity of light in N^{a}. If the z-axis is parallel to the elevator's acceleration $\boldsymbol{a}$, the spacetime metric in N^{a} has the form [41, p. 173]

$$\mathrm{d}s^2 = \left(1 + \frac{az}{c^2}\right)^2 c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 . \tag{8.3}$$

Note that, due to the existence of a horizon at $z = -c^2/a$ [41, pp. 169, 172–173], there are constraints on the size of non-inertial reference frames (accelerated or at rest in a parallel gravitational field) as represented by the metric (8.3). If the origin of N^{a} is changed, say to $z_B = 0$ (see Fig. 8.1), the horizon moves to $z = -c^2/a - |z_B|$.

As light propagates along null geodesics, with $\mathrm{d}s^2 = 0$, the coordinate velocity of light along the z-axis at a point z in N^{a} is

$$c^{\mathrm{a}}(z) = \pm c\left(1 + \frac{az}{c^2}\right) . \tag{8.4}$$

The $+$ and $-$ signs are for light propagating along or against z, respectively. Therefore, the coordinate velocity of light at a point z is locally isotropic in the z direction. It is clear that $c^{\mathrm{a}}(z)$ cannot become negative due to the constraints on the size of non-inertial frames, which ensure that $|z| < c^2/a$ [41, pp. 169, 172].

As the coordinate velocity $c^{\mathrm{a}}(z)$ is continuous on the interval $[z_A, z_B]$, one can calculate the average coordinate velocity between A and B in Fig. 8.1:

$$c^{\mathrm{a}}_{AB} = \frac{1}{z_B - z_A}\int_{z_A}^{z_B} c^{\mathrm{a}}\left(z\right)\mathrm{d}z = c\left(1 + \frac{az_B}{c^2} + \frac{ar}{2c^2}\right) , \tag{8.5}$$

where we have taken into account the fact that $z_A = z_B + r$. When the coordinate origin is at point B ($z_B = 0$), the expression (8.5) coincides with (8.1). In the same way,

$$c^{\mathrm{a}}_{BC} = c\left(1 + \frac{az_B}{c^2} - \frac{ar}{2c^2}\right) , \tag{8.6}$$

where $z_C = z_B - r$. For $z_B = 0$, (8.6) coincides with (8.2).

The coordinate velocity of light $c^{\mathrm{a}}(z)$ is also continuous on the interval $[t_A, t_B]$, but in order to calculate c^{a}_{AB} by taking an average of the velocity of light over the time of its propagation from A to B, we should find the dependence of z on t. From (8.3), we can write (for $\mathrm{d}s^2 = 0$):

$$\mathrm{d}z = c\left(1 + \frac{az}{c^2}\right)\mathrm{d}t .$$

By integrating and keeping only the terms proportional to c^{-2}, we find that $z = ct$, which shows that $c^{\mathrm{a}}(z)$ is also linear in t (to within terms proportional to c^{-2}):

$$c^{\mathrm{a}}(t) = \pm c\left(1 + \frac{at}{c}\right) .$$

Therefore, for the average coordinate velocity of light between points A and B, we have

$$\begin{aligned} c^{\mathrm{a}}_{AB} &= \frac{1}{t_B - t_A}\int_{t_A}^{t_B} c^{\mathrm{a}}(z)\,\mathrm{d}t = \frac{1}{t_B - t_A}\int_{t_A}^{t_B} c\left(1 + \frac{az}{c^2}\right)\mathrm{d}t \\ &= \frac{1}{t_B - t_A}\int_{t_A}^{t_B} c\left(1 + \frac{at}{c}\right)\mathrm{d}t = c\left(1 + \frac{az_B}{c^2} + \frac{ar}{2c^2}\right) , \end{aligned} \tag{8.7}$$

where the magnitude of $c^{\mathrm{a}}(z)$ has been used, together with $z_A = z_B + r$, $z_A = ct_A$ and $z_B = ct_B$. As expected, this expression coincides with (8.5), and for $z_B = 0$, it is equal to (8.1).

The fact that $c^{\mathrm{a}}(z)$ is linear in both z and t (to within terms $\propto c^{-2}$) makes it possible to calculate the average coordinate velocity of light propagating between A and B (see Fig. 8.1) by using the values of $c^{\mathrm{a}}(z)$ only at the end points A and B:

$$c^{\mathrm{a}}_{AB} = \frac{1}{2}\left(c^{\mathrm{a}}_A + c^{\mathrm{a}}_B\right) = \frac{1}{2}\left[c\left(1 + \frac{az_A}{c^2}\right) + c\left(1 + \frac{az_B}{c^2}\right)\right] ,$$

and since $z_A = z_B + r$,

$$c^{\mathrm{a}}_{AB} = c\left(1 + \frac{az_B}{c^2} + \frac{ar}{2c^2}\right) .$$

This expression coincides with the expressions for c^{a}_{AB} in (8.5) and (8.7).

The average coordinate velocities (8.5) and (8.6) correctly describe the propagation of light in N^{a}, yielding the right expression $\delta = ar^2/2c^2$ (see Fig. 8.1). It should be stressed that, without these average coordinate velocities, the fact that the light rays emitted from A and C arrive not at B, but at B' cannot be explained.

As a coordinate velocity, the average coordinate velocity of light is not determined with respect to a specific point and depends on the choice of the coordinate origin. Moreover, it is the same for light propagating from A to B and for light travelling in the opposite direction, i.e., $c^{\mathrm{a}}_{AB} = c^{\mathrm{a}}_{BA}$. Therefore, like the coordinate velocity (8.4), the average coordinate velocity is also isotropic but only in the sense that the average light velocity between two points is the same in both directions. As can be seen from (8.5) and (8.6), the average coordinate velocity of light between different pairs of points, whose points are the same distance apart, is different, and in this sense it is anisotropic. As a result, as shown in Fig. 8.1, the light ray emitted at A arrives at B before the light ray emitted at C.

In a non-inertial reference frame N^{g} associated with an elevator supported in a *parallel* gravitational field, where the metric is [41, p. 1056]

$$\mathrm{d}s^2 = \left(1 + \frac{2gz}{c^2}\right) c^2 \mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \,, \tag{8.8}$$

the expressions for the average coordinate velocity of light between A and B and between B and C, respectively, are

$$c^{\mathrm{g}}_{AB} = c\left(1 + \frac{gz_B}{c^2} + \frac{gr}{2c^2}\right) \tag{8.9}$$

and

$$c^{\mathrm{g}}_{BC} = c\left(1 + \frac{gz_B}{c^2} - \frac{gr}{2c^2}\right) \,.$$

8.4 Average Proper Velocity of Light

The average coordinate velocity of light explains the propagation of light in the Einstein elevator and in non-inertial reference frames in general, but cannot be used in a situation where the average light velocity between two points (say a source and an observation point) is determined *with respect to one of the points*. For instance, such situations occur in the Shapiro time delay and the situations discussed in Chap. 9. As the local velocity of light is c, the average velocity of light between a source and an observation point depends on which of

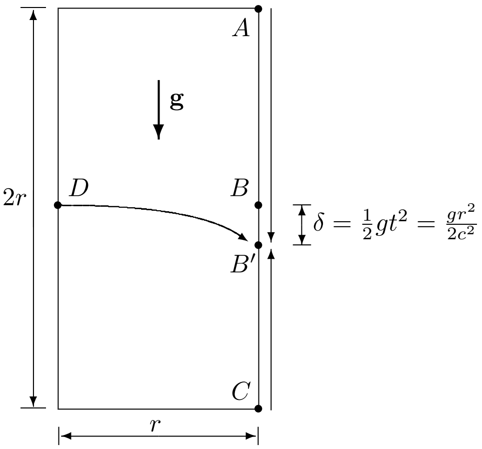


Fig. 8.2. Propagation of light in the Einstein elevator at rest in a parallel gravitational field

the two points is regarded as a reference point with respect to which the average velocity is determined. (At the reference point, the local velocity of light is always c.) The dependence of the average velocity on which point is chosen as a reference point demonstrates that that velocity is anisotropic. This anisotropic velocity can be regarded as an average *proper* velocity of light, since it is determined with respect to a given point and its calculation therefore involves the proper time at that point. It is also defined in terms of the proper distance as determined by an observer at the same point in the case of a parallel gravitational field.

Let us determine the average proper velocity of light in a non-inertial reference frame N^{g} associated with an elevator at rest in a parallel gravitational field of strength g. Consider a light source at point B (Fig. 8.2).

To calculate the average proper velocity of light which originates from B and is observed at A (that is, as seen from A), we have to determine the initial velocity of a light signal at B and its final velocity at A, both with respect to A. As the local velocity of light is c, the final velocity of the light signal determined at A is obviously c. Noting that, in a parallel gravitational field, proper and coordinate distances are the same [68], we can determine the initial velocity of the light signal at B as seen from A:

$$c_B^{\mathrm{g}} = \frac{\mathrm{d}z_B}{\mathrm{d}\tau_A} = \frac{\mathrm{d}z_B}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau_A} \ ,$$

where $\mathrm{d}z_B/\mathrm{d}t = c^{\mathrm{g}}(z_B)$ is the coordinate velocity of light at B,

$$c^{\mathrm{g}}(z_B) = c\left(1 + \frac{gz_B}{c^2}\right) ,$$

and $\mathrm{d}\tau_A = \mathrm{d}s_A/c$ is the proper time for an observer with constant spatial coordinates at A,

$$\mathrm{d}\tau_A = \left(1 + \frac{gz_A}{c^2}\right)\mathrm{d}t .$$

As $z_A = z_B + r$ and $gz_A/c^2 < 1$ (since for any value of z in N^{g}, there is a restriction $|z| < c^2/g$), for the coordinate time $\mathrm{d}t$, we have (to within terms $\propto c^{-2}$)

$$\mathrm{d}t \approx \left(1 - \frac{gz_A}{c^2}\right)\mathrm{d}\tau_A = \left(1 - \frac{gz_B}{c^2} - \frac{gr}{c^2}\right)\mathrm{d}\tau_A .$$

Then for the initial velocity c^{g}_B at B as seen from A, we obtain

$$c^{\mathrm{g}}_B = c\left(1 + \frac{gz_B}{c^2}\right)\left(1 - \frac{gz_B}{c^2} - \frac{gr}{c^2}\right) ,$$

or, keeping only the terms proportional to c^{-2},

$$c^{\mathrm{g}}_B = c\left(1 - \frac{gr}{c^2}\right) . \tag{8.10}$$

Therefore an observer at A will determine that a light signal is emitted at B with the velocity (8.10) and during the time of its journey toward A (away from the Earth's surface) will *accelerate* with an acceleration g and will arrive at A with a velocity exactly equal to c.

For the average proper velocity $\bar{c}^{\mathrm{g}}_{BA} = (1/2)(c^{\mathrm{g}}_B + c)$ of light propagating from B to A as seen from A, we have

$$\bar{c}^{\mathrm{g}}_{BA}(\text{as seen from } A) = c\left(1 - \frac{gr}{2c^2}\right) . \tag{8.11}$$

As the local velocity of light at A (measured at A) is c, it follows that if a light signal propagates from A toward B, its initial velocity at A is c, its final velocity at B is (8.10) and therefore, as seen from A, it is subjected to a negative acceleration g and will *slow down* as it 'falls' in the Earth's gravitational field. This shows that the average proper speed $\bar{c}^{\mathrm{g}}_{AB}$ (as seen from A) of a light signal emitted at A with the initial velocity c and arriving at B with the final velocity (8.10) will be equal to the average proper speed $\bar{c}^{\mathrm{g}}_{BA}$ (as seen from A) of a light signal propagating from B toward A. Thus, as seen from A, the back and forth average proper speeds of light travelling between A and B are the *same*.

Now let us determine the average proper velocity of light between B and A with respect to point B. A light signal emitted at B as seen from B will have an initial (local) velocity c there. The final velocity of the signal at A as seen from B will be

$$c_A^{\mathrm{g}} = \frac{\mathrm{d}z_A}{\mathrm{d}\tau_B} = \frac{\mathrm{d}z_A}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau_B} ,$$

where $\mathrm{d}z_A/\mathrm{d}t = c^{\mathrm{g}}(z_A)$ is the coordinate velocity of light at A,

$$c^{\mathrm{g}}(z_A) = c\left(1 + \frac{gz_A}{c^2}\right) ,$$

and $\mathrm{d}\tau_B$ is the proper time at B,

$$\mathrm{d}\tau_B = \left(1 + \frac{gz_B}{c^2}\right)\mathrm{d}t .$$

Then as $z_A = z_B + r$, we obtain for the velocity of light at A, as determined from B,

$$c_A^{\mathrm{g}} = c\left(1 + \frac{gr}{c^2}\right) . \qquad (8.12)$$

Using (8.12), the average proper velocity of light propagating from B to A as determined from B becomes

$$\bar{c}_{BA}^{\mathrm{g}}(\text{as seen from } B) = c\left(1 + \frac{gr}{2c^2}\right) . \qquad (8.13)$$

If a light signal propagates from A to B, its average proper speed $\bar{c}_{AB}^{\mathrm{g}}$ (as seen from B) will be equal to $\bar{c}_{BA}^{\mathrm{g}}$ (as seen from B) – the average proper speed of light propagating from B to A. This demonstrates that, for an observer at B, a light signal emitted from B with velocity c will *accelerate* toward A with an acceleration g and will arrive there with the final velocity (8.12). As determined by the B-observer, a light signal emitted from A with initial velocity (8.12) will be *slowing down* (with $-g$) as it 'falls' in the Earth's gravitational field and will arrive at B with a final velocity exactly equal to c. Therefore an observer at B will agree with an observer at A that a light signal will *accelerate* with an acceleration g on its way from B to A and will *decelerate* while 'falling' in the Earth's gravitational field during its propagation from A to B, but disagree on the velocity of light at the points A and B.

Comparing (8.11) and (8.13) demonstrates that the two average proper velocities between the same points A and B are not equal and depend on where they are measured from. As expected, the fact that the local velocity of light at the reference point is c makes the average proper velocity between two points dependent on where the reference point is. A consequence from here is that the Shapiro time delay does not always mean that it takes more time for light to travel a given distance in a parallel gravitational field than the time needed in flat spacetime.

In the case of a parallel gravitational field, the Shapiro time effect for a round trip of a light signal propagating between A and B determined from point A will indeed be a delay effect:

$$\Delta\tau_A = \frac{2r}{c\,(1 - gr/2c^2)} \approx \Delta t_{\text{flat}}\left(1 + \frac{gr}{2c^2}\right) ,$$

where $\Delta t_{\text{flat}} = 2r/c$ is the time for the round trip of light between A and B in flat spacetime. However, an observer at B will determine that it takes less time for a light signal to complete the round trip between A and B:

$$\Delta\tau_B = \frac{2r}{c\,(1 + gr/2c^2)} \approx \Delta t_{\text{flat}}\left(1 - \frac{gr}{2c^2}\right) .$$

On the other hand, in the Schwarzschild metric, the Shapiro effect is always a delay effect since the average proper speed of light in that metric is always smaller than c, as shown in Sect. 8.5.

The average proper velocity of light between A and B can also be obtained by using the average coordinate velocity of light (8.9) between the same points:

$$c^{\text{g}}_{AB} \equiv \frac{r}{\Delta t} = c\left(1 + \frac{gz_B}{c^2} + \frac{gr}{2c^2}\right) .$$

Let us calculate the average proper velocity of light propagating between A and B, as determined from point A. This means that we will use A's proper time $\Delta\tau_A = (1 + gz_A/c^2)\,\Delta t$:

$$\bar{c}^{\text{g}}_{AB}(\text{as seen from } A) = \frac{r}{\Delta\tau_A} = \frac{r}{\Delta t}\frac{\Delta t}{\Delta\tau_A} .$$

Noting that $r/\Delta t$ is the average coordinate velocity (8.9) and also that $z_A = z_B + r$, we have (to within terms $\propto c^{-2}$)

$$\bar{c}^{\text{g}}_{AB}(\text{as seen from } A) \approx c\left(1 + \frac{gz_B}{c^2} + \frac{gr}{2c^2}\right)\left(1 - \frac{gz_A}{c^2}\right) \approx c\left(1 - \frac{gr}{2c^2}\right) ,$$

which coincides with (8.11).

The calculation of the average proper velocity of light propagating between A and B but as seen from B yields the same expression as (8.13):

$$\begin{aligned}\bar{c}^{\mathrm{g}}_{AB}(\text{as seen from } B) &= \frac{r}{\Delta\tau_B} = \frac{r}{\Delta t}\frac{\Delta t}{\Delta\tau_B} \\ &\approx c\left(1+\frac{gz_B}{c^2}+\frac{gr}{2c^2}\right)\left(1-\frac{gz_B}{c^2}\right) \\ &\approx c\left(1+\frac{gr}{2c^2}\right) .\end{aligned}$$

Clearly, from (8.11) and (8.13), the average proper velocity of light emitted from a common source and determined at different points around the source is anisotropic in N^{g} – if the observation point is above the light source the average proper velocity of light is slightly smaller than c and smaller than the average proper velocity as determined from an observation point below the source. If an observer at point B (see Fig. 8.2) determines the average proper velocities of light coming from A and C, he will find that they are also anisotropic – the average proper velocity of light coming from A is greater than that emitted at C, and therefore the light from A will arrive at B *before* the light from C (provided that the two light signals from A and C are emitted simultaneously in N^{g}). However, if the observer at B (Fig. 8.2) determines the back and forth average proper speeds of light propagating between A and B, he finds that they are the same (the back and forth average proper speeds of light between B and C are also the same).

Let us now obtain the average proper velocity of light in vectorial form, which will be needed in Chaps. 9 and 10. Consider a light source at point B. Let the light emitted from B be observed at different points lying on a sphere of radius r and center B.

To calculate the average proper velocity of light originating from B and observed at a point P on the sphere (that is, as seen from P) we have to determine the initial velocity of a light signal at B and its final velocity at P, both with respect to P. As the local velocity of light is c, the final velocity of the light signal determined at P is obviously c. As proper and coordinate distances are the same in a parallel gravitational field, we can determine the initial velocity of the light signal at B as seen from P:

$$c^{\mathrm{g}}_B = \frac{\mathrm{d}r_B}{\mathrm{d}\tau_P} = \frac{\mathrm{d}r_B}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau_P} ,$$

where $\mathrm{d}r_B/\mathrm{d}t = c^{\mathrm{g}}(z_B)$ is the coordinate velocity of light at B. To show that the coordinate velocity of light at B is a function of z, let $\mathrm{d}r^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$. Then the interval is

$$\mathrm{d}s^2 = \left(1 + \frac{2gz}{c^2}\right) c^2 \mathrm{d}t^2 - \mathrm{d}r^2 \ ,$$

and (for $\mathrm{d}s^2 = 0$)

$$c^{\mathrm{g}}(z) \equiv \frac{\mathrm{d}r}{\mathrm{d}t} = c\left(1 + \frac{gz}{c^2}\right) \ .$$

At B the coordinate velocity is obviously

$$c^{\mathrm{g}}(z_B) = c\left(1 + \frac{gz_B}{c^2}\right) \ .$$

The proper time $\mathrm{d}\tau_P$ at P is

$$\mathrm{d}\tau_P = \left(1 + \frac{gz_P}{c^2}\right) \mathrm{d}t \ .$$

We can express z_P in terms of z_B and the distance r between B and P: $z_P = z_B - r\cos\theta$, where θ is the angle between the line of length r connecting B and P and the gravitational acceleration $\boldsymbol{g}$. As point P can be located anywhere on the sphere, we can write

$$z_P = z_B - r\cos\theta = z_B - \hat{\boldsymbol{g}} \cdot \boldsymbol{r}_P \ ,$$

where $\hat{\boldsymbol{g}}$ is a unit vector in the direction of $\boldsymbol{g}$, i.e., $\hat{\boldsymbol{g}} = \boldsymbol{g}/g$, and $\boldsymbol{r}_P$ is a position vector (with its origin at B) determining the location of P. Therefore,

$$\begin{aligned}\mathrm{d}\tau_P &= \left(1 + \frac{gz_P}{c^2}\right) \mathrm{d}t = \left(1 + \frac{gz_B}{c^2} - \frac{g\hat{\boldsymbol{g}} \cdot \boldsymbol{r}_P}{c^2}\right) \mathrm{d}t \\ &= \left(1 + \frac{gz_B}{c^2} - \frac{\boldsymbol{g} \cdot \boldsymbol{r}_P}{c^2}\right) \mathrm{d}t \ .\end{aligned}$$

As $gz_B/c^2 < 1$ and $\boldsymbol{g} \cdot \boldsymbol{r}_P/c^2 < 1$, for the coordinate time $\mathrm{d}t$ we have (to within terms $\propto c^{-2}$)

$$\mathrm{d}t \approx \left(1 - \frac{gz_B}{c^2} + \frac{\boldsymbol{g} \cdot \boldsymbol{r}_P}{c^2}\right) \mathrm{d}\tau_P \ .$$

Then for the initial velocity c_B^{g} at B as seen from P we obtain

$$c_B^{\mathrm{g}} = c\left(1 + \frac{gz_B}{c^2}\right)\left(1 - \frac{gz_B}{c^2} + \frac{\boldsymbol{g}\cdot\boldsymbol{r}_P}{c^2}\right) ,$$

or, keeping only the terms $\propto c^{-2}$,

$$c_B^{\mathrm{g}} = c\left(1 + \frac{\boldsymbol{g}\cdot\boldsymbol{r}_P}{c^2}\right) .$$

For the average proper velocity $\bar{c}_{BP}^{\mathrm{g}} = (1/2)(c_B^{\mathrm{g}} + c)$ of light propagating from B to P as seen from P, we have

$$\bar{c}_{BP}^{\mathrm{g}} = c\left(1 + \frac{\boldsymbol{g}\cdot\boldsymbol{r}_P}{2c^2}\right) ,$$

or simply

$$\bar{c}^{\mathrm{g}} = c\left(1 + \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{2c^2}\right) , \tag{8.14}$$

since $\boldsymbol{r}$ determines the point from which the average velocity is calculated.

The calculation of the average proper velocity of light in an accelerating frame N^{a} gives

$$c_{BA}^{\mathrm{a}}\,(\text{as seen from } A) = c\left(1 - \frac{ar}{2c^2}\right)$$

and

$$c_{BA}^{\mathrm{a}}\,(\text{as seen from } B) = c\left(1 + \frac{ar}{2c^2}\right) ,$$

where $a= |\boldsymbol{a}|$ is the proper acceleration of the frame. Concerning the expression for the average proper velocity of light in vectorial form, we obtain

$$\bar{c}^{\mathrm{a}} = c\left(1 - \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{2c^2}\right) . \tag{8.15}$$

Substituting $\boldsymbol{a} = -\boldsymbol{g}$ into (8.15) gives (8.14), as required by the equivalence principle. The average proper velocities of light (8.14) and (8.15) were derived *independently* and they obey the equivalence principle. This is an indication that the identical anisotropy in the propagation of light in non-inertial reference frames might have a common origin. And indeed what causes the identical anisotropic propagation of light in an accelerating reference frame N^{a} and in the frame N^{g} of an observer at rest in a gravitational field is the fact that the deformation of the worldlines of all objects that are at rest in N^{a} is *identical* to the deformation of the worldlines of the objects at rest in N^{g}. Stated another way, the worldlines of the objects at rest in N^{a} are as much

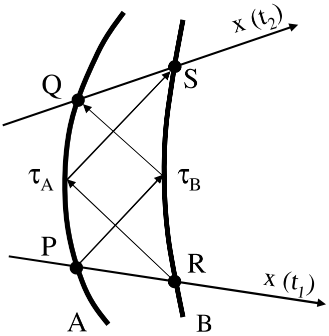


Fig. 8.3. Two observers A and B are represented by their worldlines. Observer A sends a light signal toward B at event P and receives it back at event Q, after the signal has been reflected from B. B performs the same experiment – he sends the light signal at R and receives it at S

deviated from their geodesic shapes (i.e., from their straight shapes in flat spacetime) as the worldlines of the objects that are at rest in N^{g} are deviated from their geodesic shape in curved spacetime.

To see what the shape of the worldlines of bodies in N^{a} and N^{g} has to do with the propagation of light there, consider two observers A and B at rest in N^{g}, as shown in Fig. 8.3. As we have seen in Chap. 4, the instantaneous spaces of a non-inertial reference frame corresponding to different moments of the time in that frame are not parallel to one another. In Fig. 8.3, the instantaneous spaces represented by the lines of simultaneity $x(t_1)$ and $x(t_2)$ are not parallel, which means that the proper time τ_{PQ} of observer A is not equal to the proper time τ_{RS} of observer B. It turns out that $\tau_{PQ} < \tau_{RS}$, because the worldline of A is more curved than the worldline of B.[6] The reason can only be stated here since it is beyond the scope of this book – distant objects (with respect to the direction of the frame's acceleration) that are at rest in N^{a} have different accelerations, and this means that their worldlines are curved differently.

Due to their different proper times, each of the observers determines different values of the proper velocity of light. In Fig. 8.3, at

[6] In Chap. 4, we saw that in spacetime the straight time-like worldline connecting two events is the longest distance, which means that the longest time between the two events will be measured along the straight worldline; any other, non-straight worldline connecting the same events will be shorter and therefore the time between the events measured along that worldline will be shorter. In Fig. 8.3, the worldline of A is more deformed than the worldline of B (more deviated from its straight shape) and therefore A's proper time is shorter than B's proper time.

event P, observer A sends a light signal toward B, where the signal is reflected back and after a proper time τ_{PQ} arrives at A at the event Q. B performs the same experiment and finds that it takes longer for the light signal emitted at R to return to S, since $\tau_{RS} > \tau_{PQ}$. By performing this experiment, A and B verify that $x(t_1)$ and $x(t_2)$ are lines of simultaneity in N^{a}.

Now imagine that what is depicted in Fig. 8.3 are two observers in N^{g}. Everything is exactly the same. The worldline of observer A is more deformed since it is deviated from its geodesic shape in a region where the spacetime curvature is greater (A is closer to the gravitation center). Note that the curvature of A's and B's worldlines is *not* caused by the curvature of spacetime. The worldlines are differently deformed in the sense that they are differently deviated from their normal curvature due to the spacetime curvature; that is, they are differently deviated from their geodesic shapes.

So the worldlines of A and B who are at rest in N^{a} are as much deviated from their geodesic (straight) shapes as the worldlines of the same two observers when they are at rest in N^{g} are deviated from their geodesic shapes in curved spacetime. It is this identical deformation of the observers' worldlines that gives rise to the identical anisotropic propagation of light in N^{a} and N^{g}. We will see in the next two chapters that the anisotropic propagation of light causes identical electromagnetic phenomena in N^{a} and N^{g}. If mechanical experiments are carried out in N^{a} and N^{g}, they will have the same outcomes due to the same deformation of the worldlines of the objects involved in these experiments in N^{a} and N^{g}.

Up to now we have discussed the anisotropic propagation of electromagnetic interactions in N^{a} and N^{g}. But what about the other fundamental interactions? As the carriers of the strong interactions, gluons, propagate at the speed of light, their average anisotropic velocities in N^{a} and N^{g} are

$$\bar{c}^{\mathrm{a}}_{\mathrm{S}} = c\left(1 - \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{2c^2}\right) ,$$

$$\bar{c}^{\mathrm{g}}_{\mathrm{S}} = c\left(1 + \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{2c^2}\right) .$$

The carriers of the weak interactions have nonzero rest masses and therefore propagate at a lower velocity. However, their anisotropic average proper velocities can be derived in the same way that the average proper velocity of light was derived.

As the identical anisotropic propagation of the electromagnetic, strong, and weak interactions in N^{a} and N^{g} is caused by the equal deformation of the worldlines of the particles involved in these interactions in N^{a} and N^{g}, it appears that it is this equal deformation that lies behind the equivalence principle.

8.5 Shapiro Time Delay

Although it is recognized that the retardation of light (the Shapiro time delay) is caused by the reduced speed of light in a gravitational field [57, pp. 196, 197], an expression for the average velocity of light has not been derived so far. Now we shall see that the introduction of an average proper velocity of light makes it possible for this effect to be calculated by using this velocity. It is the average proper velocity of light that is needed in the Shapiro time delay, since the time measured in this effect is the proper time at a given point.

We shall consider the treatment of the Shapiro time delay in [57, Sect. 4.4]. A light (in fact, a radio) signal is emitted from the Earth (at $z_1 < 0$) and propagates in the gravitational field of the Sun, before being reflected by a target planet (at $z_2 > 0$) and travelling back to Earth. The path of the light signal (parallel to the z axis) is approximated by a straight line [57, p. 196]. The distance between this line and the Sun (along the x axis) is b. The total proper time from the emission of the light signal to its arrival back on Earth is [57, pp. 197, 198]:

$$\Delta\tau = 2\left(1 - \frac{2GM_{\odot}}{c^2\sqrt{z_1^2+b^2}}\right)\left(\frac{z_2+|z_1|}{c} + \frac{2GM_{\odot}}{c^3}\ln\frac{\sqrt{z_2^2+b^2}+z_2}{\sqrt{z_1^2+b^2}-|z_1|}\right) . \tag{8.16}$$

As the approximate distance between the Earth (at $z_1 < 0$) and the target planet (at $z_2 > 0$) is $z_2 + |z_1|$, we can define the average proper velocity of a light signal travelling that distance as determined on Earth:

$$\begin{aligned}\bar{c}^{\mathrm{g}}_{z_1z_2} &= \frac{z_2+|z_1|}{\Delta\tau_{\mathrm{Earth}}} = \frac{z_2+|z_1|}{\Delta t}\frac{\Delta t}{\Delta\tau_{\mathrm{Earth}}} \\ &= c^{\mathrm{g}}_{z_1z_2}\frac{1}{1-\dfrac{2GM_{\odot}}{c^2\sqrt{z_1^2+b^2}}} ,\end{aligned} \tag{8.17}$$

where $c^{\mathrm{g}}_{z_1 z_2} = (z_2 + |z_1|) / \Delta t$ is the average coordinate velocity of light and

$$\Delta \tau_{\mathrm{Earth}} = \left(1 - \frac{2GM_\odot}{c^2 \sqrt{z_1^2 + b^2}} \right) \Delta t$$

is the proper time as measured on Earth and obtained from the Schwarzschild metric (neglecting the effect of the Earth's gravitational field).

We have seen in Sect. 8.3 that the average coordinate velocity $c^{\mathrm{g}}_{z_1 z_2}$ can be calculated as an average either over time or over distance, so

$$c^{\mathrm{g}}_{z_1 z_2} = \frac{1}{z_2 + |z_1|} \int_{z_1}^{z_2} c'(z) \mathrm{d}z \ ,$$

where

$$c'(z) = c \left(1 - \frac{2GM_\odot}{c^2 \sqrt{z^2 + b^2}} \right)$$

is the coordinate velocity of light at a point in the case of the Schwarzschild metric. Then

$$\begin{aligned} c^{\mathrm{g}}_{z_1 z_2} &= \frac{c}{z_2 + |z_1|} \int_{z_1}^{z_2} \left(1 - \frac{2GM_\odot}{c^2 \sqrt{z^2 + b^2}} \right) \mathrm{d}z \\ &= \frac{c}{z_2 + |z_1|} \left(z_2 + |z_1| - \frac{2GM_\odot}{c^2} \ln \frac{\sqrt{z_2^2 + b^2} + z_2}{\sqrt{z_1^2 + b^2} - |z_1|} \right) \\ &= c \left[1 - \frac{2GM_\odot}{c^2 (z_2 + |z_1|)} \ln \frac{\sqrt{z_2^2 + b^2} + z_2}{\sqrt{z_1^2 + b^2} - |z_1|} \right] \ . \end{aligned}$$

By substituting this expression for the average *coordinate* velocity of light in (8.17), we can obtain the average *proper* velocity of light in the Schwarzschild metric as seen from Earth:

$$\bar{c}^{\mathrm{g}}_{z_1 z_2} = \frac{c}{1 - \dfrac{2GM_\odot}{c^2 \sqrt{z_1^2 + b^2}}} \left[1 - \frac{2GM_\odot}{c^2 (z_2 + |z_1|)} \ln \frac{\sqrt{z_2^2 + b^2} + z_2}{\sqrt{z_1^2 + b^2} - |z_1|} \right]$$

or

$$\bar{c}^{\mathrm{g}}_{z_1 z_2} \approx \left[1 + \frac{2GM_\odot}{c^2 \sqrt{z_1^2 + b^2}} - \frac{2GM_\odot}{c^2 (z_2 + |z_1|)} \ln \frac{\sqrt{z_2^2 + b^2} + z_2}{\sqrt{z_1^2 + b^2} - |z_1|} \right] \ .$$

For the total proper time

$$\Delta\tau = \frac{2(z_2 + |z_1|)}{\overline{c}^{\mathrm{g}}_{z_1 z_2}(\text{as seen from Earth})} ,$$

from the emission of the light signal to its arrival back on Earth, we have

$$\Delta\tau = \frac{2(z_2 + |z_1|)\left(1 - \dfrac{2GM_\odot}{c^2\sqrt{z_1^2 + b^2}}\right)}{c\left[1 - \dfrac{2GM_\odot}{c^2(z_2 + |z_1|)} \ln \dfrac{\sqrt{z_2^2 + b^2} + z_2}{\sqrt{z_1^2 + b^2} - |z_1|}\right]}$$

$$\approx 2\left(1 - \frac{2GM_\odot}{c^2\sqrt{z_1^2 + b^2}}\right)\left(\frac{z_2 + |z_1|}{c} + \frac{2GM_\odot}{c^3} \ln \frac{\sqrt{z_2^2 + b^2} + z_2}{\sqrt{z_1^2 + b^2} - |z_1|}\right) ,$$

and (8.16) is recovered. The total proper time can also be written (to within terms proportional to c^{-3}) as

$$\Delta\tau \approx 2\left[\frac{z_2 + |z_1|}{c} - \frac{2GM_\odot\,(z_2 + |z_1|)}{c^3\sqrt{z_1^2 + b^2}} + \frac{2GM_\odot}{c^3} \ln \frac{\sqrt{z_2^2 + b^2} + z_2}{\sqrt{z_1^2 + b^2} - |z_1|}\right] .$$

8.6 On the Gravitational Redshift

The very existence of the gravitational redshift is now regarded as a manifestation of the curvature of spacetime [41, p. 189]. It is believed that the redshift experiment cannot be observed in the flat geometry of Minkowski spacetime since the proper times τ_{bot} and τ_{top} of a bottom experimenter (who emits light signals) and a top observer (who receives the signals) cannot be different there [41, p. 189]: "One would again conclude, if flat Minkowski geometry were valid, that $\tau_{\text{bot}} = \tau_{\text{top}}$, thus contradicting the observed redshift experiment."

In fact, the gravitational redshift does not prove the curvature of spacetime since it exists in an accelerating reference frame in flat spacetime and in a parallel (homogeneous) gravitational field where the Riemann tensor is zero and therefore there is no spacetime curvature. What the redshift effect definitely proves is that the worldlines of the non-inertial bottom experimenter and the non-inertial top observer are differently curved in flat spacetime; if the non-inertial experimenter and observer are in curved spacetime, their worldlines are differently curved there. As shown in Fig. 8.4, the different curvature of the worldlines of the bottom experimenter and the top observer explains why their proper times τ_{bot} and τ_{top} are different. So the redshift effect is caused by the fact that the two worldlines are differently curved. But two worldlines can be differently curved in a flat or curved spacetime, which means that the gravitational redshift is not caused by the curvature of spacetime.

It should be stressed that the curvature of the experimenter's and the observer's worldlines is *not* caused by the curvature of spacetime. As discussed in Sect. 8.4 the worldlines are differently curved there in the sense that they are differently deviated from their normal curvature due to the curvature of spacetime; that is, they are differently deviated from their geodesic shape. The worldline of the bottom experimenter is more deformed since it is deviated from its geodesic shape in a region where the spacetime curvature is greater (the experimenter is closer to the gravitation center). The worldline of the bottom experimenter in an accelerating reference frame is also more deformed since the acceleration of the different points of the frame is not the same in the theory of relativity.

In an accelerating reference frame, the redshift effect is often derived and explained in terms of the Doppler effect. Such an explanation, however, does not reveal the true origin of this effect. First, that explanation involves an inertial reference frame, whereas it should be done in the accelerating reference frame itself. The reason is that, by the equivalence principle, what is observed in an accelerating reference frame with proper acceleration a should also be observed in the non-inertial frame of an observer supported in a gravitational field of strength $g = a$; therefore the equivalence principle relates the gravitational redshift in both *non-inertial* frames. Second, the experimenter and the observer are *not* in relative motion, which shows that the use of the Doppler effect to describe the redshift in the accelerating frame is rather puzzling; the different curvature of the worldlines of the experimenter and the observer explains why the Doppler effect description works,[7] but does not justify it, since the fact is that the experimenter and the observer are at rest with respect to each other. As discussed

[7] As the worldlines of the experimenter and the observer are *differently* curved (not 'parallel') as shown in Fig. 8.4, it *appears* that the experimenter and the observer are in relative motion. This makes the situation similar to the standard application of the Doppler effect involving an experimenter and an observer whose worldlines are not parallel.

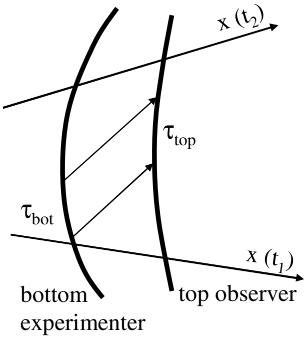


Fig. 8.4. A bottom experimenter (who emits light signals) and a top observer (who receives the signals) are at rest with respect to each other. Due to the different curvature of their worldlines, the time between the emission of two light signals $\tau_{\rm bot}$ is different from the time between the arrival of the signals $\tau_{\rm top}$. The different curvature of the worldlines of the experimenter and the observer does not imply that they are not at rest with respect to each other. As at any moment of their proper times they have different *non-parallel* instantaneous three-dimensional spaces (belonging to the instantaneous inertial reference frames at those moments), the distance between them remains constant. For instance, the spaces $x(t_1)$ and $x(t_2)$ intersect the two worldlines in such a way that the spatial distance between them is the same

above, it is the different curvature of the worldlines of the experimenter and the observer that causes the redshift effect, no matter whether that different deviation of the worldlines from their geodesic shape happens in flat or curved spacetime. If the different deformation of the worldlines of the experimenter and the observer in the gravitational case were caused by the curvature of spacetime, one might be tempted to say that the gravitational redshift is a manifestation of the spacetime curvature, whereas the redshift in the accelerating frame should be explained in terms of the Doppler effect, since there is no spacetime curvature there. However, as we have seen, this is not the case since the redshift has the *same* origin in both non-inertial reference frames (accelerating and associated with an observer at rest in a gravitational field).

An explanation of the gravitational redshift that works equally well in an accelerating reference frame and the reference frame of an observer at rest in a gravitational field is in terms of the anisotropic velocity of light in those frames. This explanation makes it more transparent that the gravitational redshift is not caused by the curvature of spacetime. The anisotropic velocity of light in both frames is caused by the curved (non-geodesic) worldlines of the experimenter and the observer. What matters is the curvature of these worldlines, not whether they are curved in flat or curved spacetime as discussed above. In this section we will derive the gravitational redshift effect only in the non-inertial reference frame of an observer at rest in a gravitational field; the derivation of the redshift effect in an accelerating reference frame is easily obtainable. By taking into account the anisotropic propagation of light in non-inertial reference frames, we will be able to:

- show that it is the frequency of a photon that is constant in the gravitational redshift,
- describe the mechanism responsible for the change of its wavelength.

It is usually assumed that both the frequency and wavelength of a photon change in the gravitational redshift, whereas its velocity remains constant. Here we will show that it is the frequency of a photon that does not change, whereas its velocity and wavelength change. It will also be shown that it is the change in the coordinate velocity of the photon along its path that leads to a change in its wavelength.

Three things should be kept in mind when dealing with the gravitational redshift:

- If two observers at different points A and B in a gravitational field determine the characteristics of a photon emitted from identical atoms placed at A and B, each observer will find that the photon characteristics – frequency, wavelength and local velocity – will have the same numerical values.
- In a *parallel* gravitational field, coordinate and proper distances coincide $\mathrm{d}x = \mathrm{d}x_A = \mathrm{d}x_B$ [68] and therefore the wavelength of a photon *at a point* is the same for all observers, i.e., $\lambda_A = \lambda_B = \lambda$.
- The local velocity of a photon *at a point* is different for different observers. (It is c only for an observer at that point.)

Consider a non-inertial frame N^{g} at rest in a *parallel* gravitational field of strength $\boldsymbol{g}$. If the z-axis is anti-parallel to the acceleration $\boldsymbol{g}$, the spacetime metric in N^{g} has the form [41, p. 1056]

$$\mathrm{d}s^2 = \left(1 + \frac{2gz}{c^2}\right) c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 , \tag{8.18}$$

whence the coordinate velocity of light at a point z in a parallel gravitational field is immediately obtained (for $\mathrm{d}s^2 = 0$) as

$$c^{\mathrm{g}} = c\left(1 + \frac{gz}{c^2}\right) . \tag{8.19}$$

Notice that in a parallel gravitational field, proper and coordinate times do not coincide (except for the proper time of an observer at infinity), whereas proper and coordinate distances are the same [68].

Consider a stationary atom at a point B in a parallel gravitational field. The atom emits a photon – a B-photon – which is observed at a point A at distance h above B. As seen at B, the B-photon is emitted with frequency $f_B = (\mathrm{d}\tau_B)^{-1}$, where $\mathrm{d}\tau_B$ is the *proper* period. As seen from A, however, the B-photon's period is $\mathrm{d}\tau_A$ and its frequency is therefore $f_B^A = (\mathrm{d}\tau_A)^{-1}$. Notice that, if an identical atom at A emits a photon, its frequency at A will be $f_A = (\mathrm{d}\tau_A)^{-1} = f_B$, which means that the corresponding periods will be (numerically) equal: $\mathrm{d}\tau_A = \mathrm{d}\tau_B$. In the case of the redshift experiment, however, when a B-photon is measured at A, $\mathrm{d}\tau_A$ and $\mathrm{d}\tau_B$ are different – $\mathrm{d}\tau_B$ is the *proper* period (measured at B) whereas $\mathrm{d}\tau_A$ is the *observed* period as measured at A. $\mathrm{d}\tau_A$ and $\mathrm{d}\tau_B$ are the proper times at A and B that correspond to the *same* coordinate time, i.e., the same coordinate period $\mathrm{d}t$:

$$\mathrm{d}\tau_A = \left(1 + \frac{gz_A}{c^2}\right)\mathrm{d}t$$

and

$$\mathrm{d}\tau_B = \left(1 + \frac{gz_B}{c^2}\right)\mathrm{d}t\ .$$

As $z_A = z_B + h$, it follows from (8.18) that the ratio of $\mathrm{d}\tau_A$ and $\mathrm{d}\tau_B$ is

$$\frac{\mathrm{d}\tau_A}{\mathrm{d}\tau_B} = \frac{(1 + gz_A/c^2)}{(1 + gz_B/c^2)} \approx 1 + \frac{gh}{c^2}\ .$$

Therefore, the *initial* frequency of the B-photon at B as seen from A will be

$$f_A = \frac{1}{\mathrm{d}\tau_A} = \frac{1}{\mathrm{d}\tau_B\,(1 + gh/c^2)} \approx f_B\left(1 - \frac{gh}{c^2}\right)\ . \tag{8.20}$$

Equation (8.20) shows that, for an observer at A, the B-photon is emitted with a reduced initial frequency $f_A < f_B$. This demonstrates that the frequency of the B-photon does not change during its journey from A to B, because its final frequency at A should also be (8.20) since $\mathrm{d}\tau_A$ is the same.

The same expression for the initial frequency of the B-photon at B as seen from A can be obtained if one makes use of the fact that, in a parallel gravitational field, proper and coordinate distance coincide. This means that the initial wavelength λ_A of the B-photon at B as seen from A is equal to the initial wavelength λ_B as measured at B, i.e., $\lambda_A = \lambda_B = \lambda$. The initial velocity of the B-photon at B as seen from A is easily calculated to be

$$c_A = \frac{\mathrm{d}z_B}{\mathrm{d}\tau_A} = \frac{\mathrm{d}z_B}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tau_A} ,$$

where and $\mathrm{d}z_B/\mathrm{d}t$ is the coordinate velocity at point B,

$$c^{\mathrm{g}} = c\left(1 + \frac{gz_B}{c^2}\right) ,$$

and

$$\mathrm{d}t = \left(1 - \frac{gz_A}{c^2}\right)\mathrm{d}\tau_A .$$

As $z_A = z_B + h$, we can write

$$c_A = c\left(1 - \frac{gh}{c^2}\right) . \tag{8.21}$$

Hence, the frequency of the B-photon at B as seen from A is

$$f_A = \frac{c_A}{\lambda} = f_B\left(1 - \frac{gh}{c^2}\right) ,$$

where $f_B = c/\lambda$.

The fact that the B-photon frequency does not change demonstrates that its energy is constant – an indication that the photon is not losing energy while moving against the gravitational field. Conversely, if an A-photon is observed at B, its constant energy will indicate that it is not gaining energy and therefore is not falling in the gravitational field. (If it were falling, its average downward speed would be greater than its upward average speed, which is not the case.)

The initial velocity of the B-photon at B, as seen from A, is given by (8.21); its final velocity at A, as seen from A, should obviously be c. The change in the photon's velocity on its way toward A also explains the mechanism responsible for the change in its wavelength. As seen from A, any wavefront moving away from the gravitational field (toward A) acquires a greater velocity as compared to the velocity of the next wavefront that follows it. Due to the speeding up of the first wavefront, the spacing between the two wavefronts increases for one period $\mathrm{d}\tau_A$ (as seen by A) by a fraction $\delta\lambda = \delta c\,\mathrm{d}\tau_A$, where

$$\delta c = c\left[1 + \frac{g\,(z + \mathrm{d}z)}{c^2}\right] - c\left(1 + \frac{gz}{c^2}\right) = c\frac{g\mathrm{d}z}{c^2}$$

is the change in the coordinate velocity over the distance dz. Then the total increase in the wavelength from B to A is

$$\Delta\lambda = \int_0^h \delta c \mathrm{d}\tau_A = c\frac{g\mathrm{d}\tau_A}{c^2}\int_0^h \mathrm{d}z = c\frac{gh}{c^2}\mathrm{d}\tau_A \ .$$

As

$$\mathrm{d}\tau_A = \mathrm{d}\tau_B\left(1 + \frac{gh}{c^2}\right) \ ,$$

we can write for $\Delta\lambda$, keeping only the terms proportional to c^{-2},

$$\Delta\lambda = c\frac{gh}{c^2}\mathrm{d}\tau_B = \lambda\frac{gh}{c^2} \ ,$$

where $c\mathrm{d}\tau_B = \lambda$ is the initial wavelength as determined at B. The final (measured) wavelength of the B-photon at A is then

$$\lambda_A = \lambda + \Delta\lambda = \lambda\left(1 + \frac{gh}{c^2}\right) \ .$$

Therefore, in the gravitational redshift, it is the velocity and wavelength of a photon that change whereas its frequency does not change.

8.7 The Sagnac Effect

The Sagnac effect can be described as follows. Two light signals emitted from a point M on the rim of a rotating disk and propagating along its rim in opposite directions will not arrive simultaneously at M. There still exist people who question special relativity and their main argument has been this effect. They claim that for an observer on the rotating disk, the speed of light is not constant – that the Galilean law of velocity addition ($c + v$ and $c - v$, where v is the orbital speed at a point on the disk rim) should be used by the rotating observer in order to explain the time difference in the arrival of the two light signals at M. Therefore, according to relativity dissidents, one can discover the absolute motion of a point on the rim of the disk.

What makes such claims even more persistent is the lack of a clear position on the issue of the speed of light in non-inertial reference frames. What special relativity states is that the speed of light is constant only in inertial reference frames – this constancy follows from the impossibility of detecting absolute uniform motion. (More precisely, it follows from the non-existence of absolute uniform motion.) Accelerated motion can be detected and for this reason the coordinate velocity of light in non-inertial reference frames is a function of the proper acceleration of the frame. The rotating disk is a non-inertial reference frame and its acceleration can be detected by different means including light signals. That is why it is not surprising that the coordinate velocity of light as determined on the disk depends on the centripetal acceleration of the disk. As we shall see below, the coordinate velocity of light calculated on the disk is *not* a manifestation of the Galilean law of velocity addition.

Consider two disks whose centers coincide. One of them is stationary, the other rotates with constant angular velocity ω. As the stationary disk can be regarded as an inertial frame, its metric is the Minkowski metric:

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \; . \qquad (8.22)$$

To write the interval $\mathrm{d}s^2$ in polar coordinates, we use the transformation

$$t = t \; , \quad x = R\cos\Phi \; , \quad y = R\sin\Phi \; , \quad z = z \; . \qquad (8.23)$$

Substituting (8.22) into (8.23), we get

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - \mathrm{d}R^2 - R^2\mathrm{d}\Phi^2 - \mathrm{d}z^2 \; . \qquad (8.24)$$

Let an observer on the rotating disk use the coordinates t, r, φ, and z. The transformation between the coordinates on the stationary and on the rotating disk is obviously

$$t = t \; , \quad R = r \; , \quad \Phi = \varphi + \omega t \; , \quad z = z \; . \qquad (8.25)$$

Time does not change in this transformation since the coordinate time on the rotating disk is given by the clock at its center and this clock is at rest with respect to the inertial stationary disk [42]. Substituting (8.25) into (8.24), we obtain the metric on the rotating disk:

$$\mathrm{d}s^2 = \left(1 - \frac{\omega^2 r^2}{c^2}\right) c^2\mathrm{d}t^2 - \mathrm{d}r^2 - r^2\mathrm{d}\varphi^2 - 2\omega r^2\mathrm{d}t\mathrm{d}\varphi - \mathrm{d}z^2 \; . \quad (8.26)$$

As light propagates along null geodesics ($\mathrm{d}s^2 = 0$), we can calculate the tangential coordinate velocity of light $c^{\varphi} \equiv r\,(\mathrm{d}\varphi/\mathrm{d}t)$ from (8.26) by taking into account the fact that $\mathrm{d}r = 0$ and $\mathrm{d}z = 0$ for light propagating on the surface of the rotating disk along its rim (of radius r). First we have to determine $\mathrm{d}\varphi/\mathrm{d}t$. From (8.26), we can write

$$r^2\left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right)^2 + 2\omega r^2\left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right) - \left(1 - \frac{\omega^2 r^2}{c^2}\right)c^2 = 0 \; .$$

The solution of this quadratic equation gives two values for $\mathrm{d}\varphi/\mathrm{d}t$, one in the direction in which φ increases $(+\varphi)$ (in the direction of rotation of the disk) and the other in the opposite direction $(-\varphi)$:

$$\left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right)^{+\varphi} = -\omega + \frac{c}{r}\,, \qquad \left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right)^{-\varphi} = -\omega - \frac{c}{r}\,.$$

Then for the tangential coordinate velocities $c^{+\varphi}$ and $c^{-\varphi}$, we obtain

$$c^{+\varphi} \equiv r\left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right)^{+\varphi} = c\left(1 - \frac{\omega r}{c}\right) \tag{8.27}$$

and

$$c^{-\varphi} \equiv r\left(\frac{\mathrm{d}\varphi}{\mathrm{d}t}\right)^{-\varphi} = -c\left(1 + \frac{\omega r}{c}\right). \tag{8.28}$$

As can be seen from (8.27) and (8.28), the tangential coordinate velocities $c^{+\varphi}$ and $c^{-\varphi}$ are *constant* for a given r, which means that (8.27) and (8.28) also represent the average coordinate velocities of light. The coordinate speed of light propagating in the direction of rotation of the disk is smaller than the coordinate speed in the opposite direction.

This fact allows an observer on the rotating disk to explain why two light signals emitted from a point M on the disk rim and propagating along the rim in opposite directions will not arrive simultaneously at M – as the coordinate speed of the light signal travelling against the disk rotation is greater than the speed of the other signal, it will arrive at M first.

The time it takes a light signal travelling along the rim of the disk in the direction of its rotation to complete one revolution is

$$\Delta t^{+\varphi} = \frac{2\pi r}{c^{+\varphi}} = \frac{2\pi r}{c\,(1 - \omega r/c)} = \frac{2\pi r}{c - \omega r}\,.$$

The time for the completion of one revolution by the light signal propagating in the opposite direction is

$$\Delta t^{-\varphi} = \frac{2\pi r}{|c^{-\varphi}|} = \frac{2\pi r}{c\,(1 + \omega r/c)} = \frac{2\pi r}{c + \omega r}\,.$$

The arrival of the two light signals at M is separated by the time interval

$$\delta t = \Delta t^{+\varphi} - \Delta t^{-\varphi} = \frac{4\pi\omega r^2}{c^2 - \omega^2 r^2}\,. \tag{8.29}$$

The time difference (8.29) is caused by the different coordinate speeds of light in the $+\varphi$ and $-\varphi$ directions. Here it should be specifically stressed that $c^{+\varphi}$ and $c^{-\varphi}$ are different from c owing to the accelerated motion (rotation) of the disk. In terms of the orbital velocity $v = \omega r$, it appears that the two tangential coordinate velocities can be written as a function of v, viz.,

$$c^{+\varphi} = c\left(1 - \frac{v}{c}\right) = c - v \ , \qquad c^{-\varphi} = c\left(1 + \frac{v}{c}\right) = c + v \ ,$$

which resemble the Galilean law of velocity addition. However, it is completely clear that this resemblance is misleading – due to the centripetal (normal) acceleration $a^{\mathrm{N}} = v^2/r$, the *direction* of the orbital velocity constantly changes during the rotation of the disk, which means that $c^{+\varphi}$ and $c^{-\varphi}$ depend on the normal acceleration of the disk:

$$c^{+\varphi} = c\left(1 - \frac{\sqrt{a^{\mathrm{N}} r}}{c}\right) \tag{8.30}$$

and

$$c^{-\varphi} = c\left(1 + \frac{\sqrt{a^{\mathrm{N}} r}}{c}\right) \ . \tag{8.31}$$

As expected, the expressions (8.30) and (8.31) are similar to the average coordinate velocities (8.5) and (8.6) (for $z_B = 0$) in the sense that all coordinate velocities depend on the acceleration, not the velocity.

8.8 Summary

This chapter revisited the question of the constancy of the speed of light by pointing out that it has two answers – the speed of light is constant in all inertial reference frames, but when determined in a non-inertial frame, it depends on the frame's proper acceleration. (The local velocity of light, however, is always c.) It has been shown that the complete description of the propagation of light in non-inertial frames of reference requires an average coordinate and an average proper velocity of light. The need for an average coordinate velocity was demonstrated in the case of Einstein's elevator thought experiment – to explain the fact that two light signals emitted from points A and C in Fig. 8.1 meet at B', not at B. It was also shown that an average proper velocity of light is implicitly used in the Shapiro time delay effect; when such a velocity is explicitly defined, it follows that, in the case of a parallel gravitational field, the Shapiro effect is not always a delay effect.

The Sagnac effect was also revisited by defining the coordinate velocity of light in the non-inertial frame of the rotating disk. This velocity naturally explains the fact that two light signals emitted from a point on the rim of the rotating disk and propagating along its rim in opposite directions do not arrive simultaneously at the same point.

9 Calculating the Electric Field of a Charge in a Non-Inertial Reference Frame

The usual way to calculate the electric field of a charge at rest in a non-inertial reference frame N (accelerating N^{a} or supported in a gravitational field N^{g}) is to transform the field from an inertial reference frame I to N (see for instance [77]). The reason is that the *direct* calculation of the potential and electric field of a charge in N turns out to be problematic. However, a direct calculation of these quantities in N^{a} and N^{g} is essential for a rigorous application of the equivalence principle, which requires that expressions for physical quantities, calculated in N^{a} and N^{g}, be compared.[1]

9.1 Calculating the Potential of a Charge in a Non-Inertial Reference Frame

In 1921 Fermi [71] studied the nature of the force acting on a charge at rest in a gravitational field of strength $\boldsymbol{g}$ in the framework of general relativity and the classical electron theory. He derived the potential

$$\varphi_{\mathrm{F}}^{\mathrm{g}} = \frac{e}{4\pi\epsilon_0 r}\left(1 - \frac{1}{2}\frac{gz}{c^2}\right) , \tag{9.1}$$

where $g = |\boldsymbol{g}|$, in the non-inertial reference frame N^{g} in which the charge e is at rest. However, this contains a factor of 1/2 in the brackets which leads to a contradiction with the principle of equivalence. To see this, let us calculate the electric field from this potential [71]:

$$\boldsymbol{E}_{\mathrm{F}}^{\mathrm{g}} = -\boldsymbol{\nabla}\varphi_{\mathrm{F}}^{\mathrm{g}} = \frac{e}{4\pi\epsilon_0}\left(\frac{\boldsymbol{n}}{r^2} - \frac{\boldsymbol{g}\cdot\boldsymbol{n}}{2c^2 r}\boldsymbol{n} + \frac{\boldsymbol{g}}{2c^2 r}\right) , \tag{9.2}$$

[1] Those expressions can be obtained in the comoving inertial frame (for the case of N^{a}) and in the local inertial reference frame (for the case of N^{g}) and then compared. However, as we will see later that application (or verification) of the equivalence principle does not provide any deep physical understanding of what lies behind the equivalence of phenomena observed in N^{a} and N^{g}. Moreover, if all the phenomena are equal in N^{a} and N^{g}, then why can their effects not be described and calculated *directly* in the non-inertial reference frames themselves.

where $\boldsymbol{r}$ is the displacement vector from the point where the charge is located to the point where $\boldsymbol{E}_{\mathrm{F}}^{\mathrm{g}}$ is calculated. The equivalence principle requires that the electric field of a charge in N^{g} should be the same as that of a charge at rest in an accelerating reference frame N^{a} whose acceleration is $\boldsymbol{a} = -\boldsymbol{g}$. In other words, substituting $\boldsymbol{g} = -\boldsymbol{a}$ in (9.2) should give the electric field $\boldsymbol{E}^{\mathrm{a}}$ of the charge in N^{a}. Later we will calculate $\boldsymbol{E}^{\mathrm{a}}$ directly in N^{a}, but for the moment let us compare the electric field (9.2) with the electric field of a charge e determined in an inertial reference frame I in which the accelerating charge is instantaneously at rest (see [76, p. 664] for the case of the instantaneous field, i.e., when its velocity relative to I is zero):

$$\boldsymbol{E}^{\mathrm{a}} = \frac{e}{4\pi\epsilon_0}\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{a}}{c^2 r}\right) . \tag{9.3}$$

We can do this since the instantaneous electric field of the charge in I depends only on the charge's acceleration, not on its velocity relative to I. And since acceleration is absolute, its effect on the shape of the electric field is also absolute, which means that the shape of the field should be the same for all observers (inertial and non-inertial).

A comparison of (9.2) and (9.3) shows that the electric field of a charge supported in the Earth's gravitational field coincides with the instantaneous electric field of a charge moving with an acceleration $\boldsymbol{a} = -\boldsymbol{g}/2$. As the principle of equivalence requires that $\boldsymbol{a} = -\boldsymbol{g}$, it is evident that the 1/2 factor in the terms containing $\boldsymbol{g}$ in (9.2) is an indication that something has not been taken into account in the calculations of the potential and electric field of a charge in a non-inertial reference frame.

In the last chapter we saw that, due to the absoluteness of acceleration, the propagation of light in non-inertial reference frames is anisotropic, which enables a non-inertial observer to detect his acceleration. This fact has an immediate implication for any electromagnetic calculations in a non-inertial reference frame N, since the propagation of any disturbance in the field of a charge in N is also anisotropic. To see this let us write the potential of a charge at rest in N^{g}:

$$\varphi^{\mathrm{g}} = \frac{1}{4\pi\epsilon_0}\frac{\rho V^{\mathrm{g}}}{r^{\mathrm{g}}} , \tag{9.4}$$

where ρ is the charge density, V^{g} is the volume of the charge, and the radius vector r^{g} is the distance from the charge to the observation point (where the potential is measured), both determined in N^{g}.

There are three quantities that can be affected by the anisotropic propagation of light (and any electromagnetic disturbances) in the expression for the potential (9.4). These are the charge, the distance r^{g}, and the volume V^{g}.

The experimental evidence tells us that the charge is Lorentz invariant (see [76, p. 554] and references therein). This fact in combination with another fact – that at *every* instant all physical quantities determined in a non-inertial reference frame N^{a} or N^{g} are equal to those determined in the comoving or local inertial reference frame I – show that the charge e^{g} measured in the frame N^{g} (supported in the Earth's gravitational field) is equal to the charge e measured in the local inertial frame I. We will return to this point a little later.

The radius vector r^{g} is affected by the anisotropic propagation of light in N^{g} for the following reason. Assume that the charge suddenly changes its position. At a given moment of the coordinate time t in N^{g}, the disturbance in the field of the charge will reach points around the charge, determined by the radius vector $\boldsymbol{r}^{\mathrm{g}}$, which do *not* lie on a sphere whose center is located at the point where the charge is; those points will rather form a distorted sphere. This is so since the average proper velocity of light (8.14) is different in different directions. By expressing r^{g} as $r^{\mathrm{g}} = \bar{c}^{\mathrm{g}} t$ in N^{g} and taking into account the fact that $\boldsymbol{g} \cdot \boldsymbol{r}/2c^2 < 1$, we can write (keeping only the terms $\propto c^{-2}$)

$$(r^{\mathrm{g}})^{-1} \approx r^{-1} \left(1 - \frac{\boldsymbol{g} \cdot \boldsymbol{r}}{2c^2} \right) . \tag{9.5}$$

In fact, representing r^{g} as $r^{\mathrm{g}} = \bar{c}^{\mathrm{g}} t$ is not done specifically for the calculations in N^{g}. In the calculation of the Liénard–Wiechert potentials in an inertial reference frame, the radius vector r is also written as $r = ct$ [72, p. 416].

Substituting $(r^{\mathrm{g}})^{-1}$ into (9.4) gives the potential (9.1) obtained by Fermi. This is an indication that the volume V^{g} of a charge in N^{g} will also be affected by the anisotropic propagation of light there. And indeed, as we shall now see, the anisotropic average proper velocity of light (8.14) gives rise to a Liénard–Wiechert-like (or rather anisotropic) volume V^{g} (not coinciding with the actual volume V) which removes the 1/2 factor in (9.1) and therefore in (9.2) as well.

The origin of V^{g} is analogous to the origin of the Liénard–Wiechert volume $V^{\mathrm{LW}} = V/\left(1 - \boldsymbol{v} \cdot \boldsymbol{n}/c\right)$ of a charge moving at velocity $\boldsymbol{v}$ with respect to an inertial observer I, where $\boldsymbol{n} = \boldsymbol{r}/r$ and $\boldsymbol{r}$ is the radius vector at the retarded time [72, p. 418]. One way to explain the origin of V^{LW} is in terms of the 'information-collecting sphere' of Panofsky and

Phillips [73] used in the derivation of the Liénard–Wiechert potentials; similar concepts are employed by Griffiths [72, p. 418], Feynman [74, p. 21-10], and Schwartz [75].

A charge approaching an observation point where the potential is determined contributes more to the potential there since it “stays longer within the information-collecting sphere” [73, p. 343], which moves at the velocity of light c in I while converging toward the observation point. The greater contribution to the potential may be viewed as originating from a charge of Liénard–Wiechert volume V^{LW} that appears greater[2] than V. If the charge is receding from the observation point, the information-collecting sphere moves against the charge, the charge stays less time within it and the resulting smaller contribution to the potential can be regarded as coming from a charge whose Liénard–Wiechert volume V^{LW} appears smaller than V.

By the same argument the anisotropic volume V^{g} also appears different from V in N^{g}. Consider a charge of length l at rest in N^{g} placed along the direction of $\boldsymbol{g}$. The time for which the information-collecting sphere travelling at the average velocity of light (8.14) in N^{g} sweeps over the charge is

$$\Delta t^{\mathrm{g}} = \frac{l}{\bar{c}^{\mathrm{g}}} = \frac{l}{c\,(1 + \boldsymbol{g} \cdot \boldsymbol{r}/2c^2)} \approx \Delta t^{I}\left(1 - \frac{\boldsymbol{g} \cdot \boldsymbol{r}}{2c^2}\right) ,$$

where $\Delta t^{I} = l/c$ is the time for which the information-collecting sphere propagating at speed c sweeps over an inertial charge of the same length l in its rest frame. If the observation point where the potential is calculated is above the charge, the information-collecting sphere moves against $\boldsymbol{g}$ in N^{g}, its average velocity is smaller than c (as determined at that point) and therefore $\Delta t^{\mathrm{g}} > \Delta t^{I}$ (since $\boldsymbol{g} \cdot \boldsymbol{r} = -gr$). As a result the charge stays longer within the sphere and its contribution to the potential is greater. This is equivalent to saying that the greater contribution comes from a charge of a greater length l^{g} which for the same time Δt^{g} is swept over by an information-collecting sphere propagating at velocity c:

$$l^{\mathrm{g}} = \Delta t^{\mathrm{g}} c = l\left(1 - \frac{\boldsymbol{g} \cdot \boldsymbol{r}}{2c^2}\right) .$$

[2] The fact that the volume of the charge depends on the observation point at which the potential is determined does not mean that there is a violation of the invariance of the electric charge. For a specific discussion of this point see [73, p. 342]; for a derivation of the Liénard–Wiechert volume in the calculation of the Liénard–Wiechert potentials see [72, p. 419].

If the point where the potential is determined is below the charge, the length of the charge will appear shorter since the information-collecting sphere propagates in the direction of $\boldsymbol{g}$. This means that $\boldsymbol{g} \cdot \boldsymbol{r} = gr$ and therefore the contribution of the charge to the potential at that point will be smaller.

The anisotropic volume which corresponds to such an apparent length l^{g} is obviously

$$V^{\mathrm{g}} = V\left(1 - \frac{\boldsymbol{g} \cdot \boldsymbol{r}}{2c^2}\right) . \tag{9.6}$$

Substituting (9.5) and (9.6) into (9.4) gives the scalar potential of the charge ρV^{g} as

$$\varphi^{\mathrm{g}} = \frac{1}{4\pi\epsilon_0}\frac{\rho V^{\mathrm{g}}}{r^{\mathrm{g}}} = \frac{1}{4\pi\epsilon_0}\frac{\rho V}{r}\left(1 - \frac{\boldsymbol{g} \cdot \boldsymbol{r}}{2c^2}\right)^2 ,$$

or, if we keep only the terms proportional to c^{-2},

$$\varphi^{\mathrm{g}} = \frac{\rho V}{4\pi\epsilon_0 r}\left(1 - \frac{\boldsymbol{g} \cdot \boldsymbol{r}}{c^2}\right) . \tag{9.7}$$

As can be seen from (9.7), making use of V^{g} instead of V accounts for the 1/2 factor in (9.1). This is an indication that a complete description of electromagnetic phenomena in N^{g} is not possible without using the average proper velocity of light (8.14) and the resulting anisotropic volume (9.6).

As the effect of the anisotropic propagation of light on a volume element in a non-inertial reference frame has not been noticed so far, let us examine it in more detail. There is another reason for paying specific attention to this effect. A careful analysis of the physical origin of the Liénard–Wiechert potential (similar to that carried out in [72, p. 418], [74, p. 21-10], and [75]) leads to the fundamental question: How does a charge *update* its field (and potential)? Once this is understood, the appearance of the anisotropic volume element in non-inertial reference frames comes as no surprise.

9.2 Common Physical Origin of the Liénard–Wiechert Potentials and the Potentials of a Charge in a Non-Inertial Reference Frame

In order to understand why the anisotropic propagation of light in a non-inertial reference frame N leads to a Liénard–Wiechert-like contribution to the potential of a charge measured in N, let us first consider in more detail the origin of the Liénard–Wiechert potentials themselves. As we have seen in the last section, they originate from an apparent change in the volume of a charge as seen at the observation point [72, 74].

To better realize the physical origin of this apparent change in the volume, consider a charge $e = \rho V$ of length l and volume V. The potential of the charge is determined at an observation point P in an inertial reference frame I. When the charge is at rest in I, the potential at point P at any given moment is determined by the total charge e. If the charge is in relative motion with respect to I, however, the potential at P is not determined by the total charge. To explain this we cannot avoid the open question of how the charge *updates* the potential at a given point. We shall make the assumption that the charge updates the potential at P at every instant by constantly emitting some kind of updating signals (as we shall see, the unknown frequency of these updating signals will not affect our calculations). To come up with such a hypothesis when the Liénard–Wiechert potentials were derived could have been quite an insight. Our radical research team from the first part of the book would have loved to explore the physical origin of the Liénard–Wiechert potentials. And quite possibly they would have arrived at the hypothesis that a charge constantly emits some kind of signal in order to update its potential and field. Making the assumption of updating signals which travel at speed c is now easy. In quantum electrodynamics the field of an electric charge consists of a set of virtual photons which are constantly being emitted and absorbed by the charge. The treatment here is classical but, if this helps, the updating signals can be thought of as virtual photons.

In terms of the updating signals, the potential at P at a given moment t in I is determined by all updating signals originating from different points of the charge at different moments and reaching P at the *same* moment t. The potential of a stationary charge in I is determined by all updating signals emitted by different points of the charge during the time $\Delta t = l/c$ for which the updating signal from the rear end of the charge (with respect to P) reaches the front end of the charge. After the rear end signal and the signals from the other points of the charge have reached the front end point, the last updating signal is emitted from that point and all these signals travel toward P. They arrive *simultaneously* at point P and determine the scalar potential φ there due to the total charge e.

The potential of a charge moving toward P is not represented by the total charge because it takes more (Liénard–Wiechert) time $\Delta t_{+}^{\mathrm{LW}}$

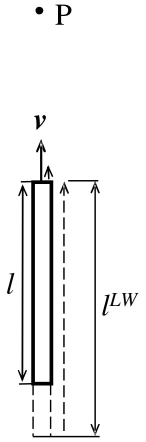


Fig. 9.1. A charge moving toward point P with velocity $\boldsymbol{v}$ relative to an inertial reference frame I. Its effective potential at P is determined by all simultaneously arriving updating signals that are emitted by the charge

for the updating signal from the rear end of the charge to reach its front end point since, during the time the rear end signal travels toward the front end point of the charge, the charge is also moving in the same direction and the signal chases the front end point (Fig. 9.1). As a result, during that time *more* updating signals are emitted, which arrive simultaneously at P. This means that the potential at P represents *not* the potential φ from the total charge e, but an *effective* potential φ_+^{LW} which is larger than φ. In this case, the time Δt_+^{LW} after which the rear end signal reaches the front end of the charge moving at speed v can be calculated directly:

$$\Delta t_+^{\mathrm{LW}} = \frac{l + v\Delta t_+^{\mathrm{LW}}}{c} \,,$$

or

$$\Delta t_+^{\mathrm{LW}} = \frac{l/c}{1 - v/c} \,.$$

If the charge is receding from P, it takes the front end updating signal less time Δt_-^{LW} to reach the rear end of the charge and fewer updating signals arrive *simultaneously* at P, which means that the potential there represents *not* the potential φ from the total charge e but an *effective* potential φ_-^{LW}, which is smaller than φ. The time Δt_-^{LW} in this case is

$$\Delta t_-^{\mathrm{LW}} = \frac{l/c}{1 + v/c} \,.$$

The vector notation for the Liénard–Wiechert time is easily obtained:

$$\Delta t^{\mathrm{LW}} = \frac{l/c}{1 - \dfrac{\boldsymbol{v} \cdot \boldsymbol{n}}{c}} \,,$$

where $\boldsymbol{n}$ is a unit vector in the direction of propagation of the updating signals (toward the observation point). The Liénard–Wiechert time Δt^{LW} during which updating signals determining the potential of the moving charge are emitted implies an apparent length l^{LW} of the charge along the line connecting the charge and the observation point which is different from its actual length l. l^{LW} is the apparent length from which all updating signals, arriving simultaneously at P, are emitted toward P – this is the distance between the rear end of the charge from which the first updating signal is emitted and the point where this signal reaches the front end of the charge moving at speed v after time Δt^{LW} from which the last updating signal is emitted (see Fig.9.1):

$$l^{\mathrm{LW}} = c\Delta t^{\mathrm{LW}} = \frac{l}{1 - \dfrac{\boldsymbol{v} \cdot \boldsymbol{n}}{c}} \,.$$

This means that the apparent volume of the charge, as seen from P, is

$$V^{\mathrm{LW}} = \frac{V}{1 - \dfrac{\boldsymbol{v} \cdot \boldsymbol{n}}{c}} \,. \tag{9.8}$$

Then the potential at point P is

$$\varphi^{\mathrm{LW}} = \frac{1}{4\pi\epsilon_0}\frac{\rho V^{\mathrm{LW}}}{r} = \frac{1}{4\pi\epsilon_0}\frac{\rho V}{r\left(1 - \dfrac{\boldsymbol{v} \cdot \boldsymbol{n}}{c}\right)} \,,$$

or

$$\varphi^{\mathrm{LW}}(r,t) = \left|\frac{1}{4\pi\epsilon_0}\frac{e}{r\left(1 - \dfrac{\boldsymbol{v} \cdot \boldsymbol{n}}{c}\right)}\right|_{\mathrm{ret}} \,, \tag{9.9}$$

which is the scalar Liénard–Wiechert potential. ρ is the density of the charge and r is the retarded position of the center of the charge. It should be emphasized that the charge density does *not* change. The charge, its density, and its actual volume V do not change. We must *presuppose* that they remain the *same* in order to derive the apparent Liénard–Wiechert volume and potentials.

After taking into account the Liénard–Wiechert volume (9.8), the vector Liénard–Wiechert potential is obviously

$$\boldsymbol{A}^{\mathrm{LW}}(r,t) = \left| \frac{e}{4\pi\epsilon_0 c^2} \frac{\boldsymbol{v}}{r\left(1 - \dfrac{\boldsymbol{v}\cdot\boldsymbol{n}}{c}\right)} \right|_{\mathrm{ret}} , \tag{9.10}$$

where r is again the retarded position of the charge.

The scalar and vector Liénard–Wiechert potentials (9.9) and (9.10) do not depend on the size of the charge. For this reason one may argue that they are valid for a point charge as well. If this were the case, the explanation of the physical origin of the Liénard–Wiechert potentials would not make sense. A point charge is a useful idealization, but many physicists doubt whether such a thing exists in nature. The strongest argument comes from quantum mechanics – recall the discussion of the dipole moment of the hydrogen atom in Chap. 6. The fact that the potentials (9.9) and (9.10) do not depend on the size of the charge simply means that they hold for charges of *any* size.

We are now in a position to derive the expression for the anisotropic volume element in a more physical way; the derivation in the previous section was based on the rather artificial concept of 'information-collecting sphere'. Consider a charge $e = \rho V$ of length l at rest in N^{g}. During a given period of time all the points of the charge emit updating signals which arrive *simultaneously* at the observation point P. The mechanism determining the period of time during which all updating signals, that reach P *simultaneously*, are emitted is the same as in the case of the Liénard–Wiechert potentials. If P is at a distance r from the front end of the charge, the time for which the updating signals from the charge are emitted is different from $\Delta t = l/c$ due to the anisotropic velocity of light. If P is above the charge, the average velocity of the updating signals is smaller than c, the charge will contribute to the observation point for a longer time, and more updating signals will arrive simultaneously at P. The reason is as follows.

An updating signal RE is emitted from the rear end of the charge (point RE) and starts to propagate toward the observation point P (Fig. 9.2). Its average proper velocity c^{g}_{RE} is

$$c^{\mathrm{g}}_{RE} = c\left[1 + \frac{\boldsymbol{g}\cdot(\boldsymbol{r}+\boldsymbol{l})}{2c^2}\right] , \tag{9.11}$$

where $\boldsymbol{l}$ is a vector parallel to r with magnitude equal to the length of the charge. The average velocity c^{g}_{RE} is slightly smaller than c since

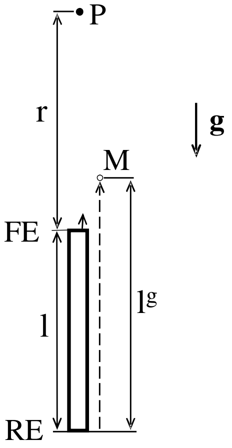


Fig. 9.2. A charge is at rest in N^{g}. Its effective potential at P is determined by all simultaneously arriving updating signals that are emitted by the charge and whose average velocities are different

$\boldsymbol{g}\cdot(\boldsymbol{r}+\boldsymbol{l}) = -g(r+l)$. During the journey of the RE signal toward P, updating signals from all points of the charge are being emitted in the direction of P as well.[3] When the RE signal reaches point M, the last signal from the front end of the charge (point FE) is emitted and all updating signals arrive *simultaneously* at P. The reason why the last signal from the charge is emitted, not when the RE signal is at point FE, but when it is at point M is that the average proper velocity of the FE signal is slightly greater than that of the RE signal:

$$c^{\mathrm{g}}_{FE} = c\left(1 + \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{2c^2}\right) . \tag{9.12}$$

Therefore, the slower RE signal should be at M in order that it arrive at P simultaneously with the faster FE signal. Since $c^{\mathrm{g}}_{RE} < c$ more signals will be emitted during the propagation of the RE signal between the points RE and M than if the RE signal moved at speed c.

The time during which all updating signals, that arrive simultaneously at P, are being emitted from the charge is

$$\begin{aligned}\Delta t^{\mathrm{g}} &= \Delta t_{RP} - \Delta t_{FP}\\ &= \frac{r+l}{c\left[1 + \dfrac{\boldsymbol{g}\cdot(\boldsymbol{r}+\boldsymbol{l})}{2c^2}\right]} - \frac{r}{c\left(1 + \dfrac{\boldsymbol{g}\cdot\boldsymbol{r}}{2c^2}\right)}\end{aligned}$$

[3] Such signals are emitted in all directions, but we are interested only in those signals which propagate toward P.

$$\approx \frac{l}{c}\left(1-\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}-\frac{\boldsymbol{g}\cdot\boldsymbol{l}}{2c^2}\right) ,$$

where Δt_{RP} is the time it takes the RE updating signal, travelling at the speed (9.11) to reach point P, and Δt_{FP} is the time it takes the last signal emitted from the front end of the charge, propagating at the speed (9.12), to reach the observation point P. The apparent length l^{g} from which all updating signals, arriving simultaneously at P, are emitted can be calculated by multiplying the time Δt^{g} and the average speed of the updating signal emitted from the rear end of the charge. The velocity (9.11) is used since, as shown in Fig. 9.2, l^{g} is the distance between the rear end of the charge, from where the first updating signal is emitted, and the position (point M) of the same signal travelling at (9.11), after the time Δt^{g}, when the last updating signal is emitted from the front end of the charge. Hence for l^{g}, we obtain

$$\begin{aligned} l^{\mathrm{g}} &= c^{\mathrm{g}}_{RE}\Delta t^{\mathrm{g}} \\ &= c\left[1+\frac{\boldsymbol{g}\cdot(\boldsymbol{r}+\boldsymbol{l})}{2c^2}\right]\frac{l}{c}\left(1-\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}-\frac{\boldsymbol{g}\cdot\boldsymbol{l}}{2c^2}\right) \\ &\approx l\left(1-\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{2c^2}\right) . \end{aligned}$$

Here we have kept only the terms proportional to c^{-2}. As expected, the apparent volume of the charge which results from l^{g} coincides with (9.6) derived in terms of the information collecting sphere:

$$V^{\mathrm{g}} = V\left(1-\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{2c^2}\right) ,$$

which leads, as we have seen, to the correct potential for a charge at rest in N^{g}:

$$\varphi^{\mathrm{g}} = \frac{\rho V}{4\pi\epsilon_0 r}\left(1-\frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right) .$$

So we now know how to calculate the potential of a charge supported in a gravitational field. But what about a charge that is moving in N^{g}? As moving charges are described by the Liénard–Wiechert potentials, we need to generalize them to the case of non-inertial reference frames. For this purpose, let us write (9.9) and (9.10) in terms of r^{g} and V^{g}:

$$\varphi^{\mathrm{LW}}_{\mathrm{g}}(r,t) = \left|\frac{1}{4\pi\epsilon_0}\frac{\rho V^{\mathrm{g}}}{r^{\mathrm{g}}\left(1-\dfrac{\boldsymbol{v}\cdot\boldsymbol{n}}{c}\right)}\right|_{\mathrm{ret}} , \tag{9.13}$$

and

$$\boldsymbol{A}_{\mathrm{g}}^{\mathrm{LW}}(r,t) = \left| \frac{\rho V^{\mathrm{g}}}{4\pi\epsilon_0 c^2} \frac{\boldsymbol{v}}{r^{\mathrm{g}}\left(1 - \dfrac{\boldsymbol{v}\cdot\boldsymbol{n}}{c}\right)} \right|_{\mathrm{ret}} . \tag{9.14}$$

The generalized Liénard–Wiechert potentials can be obtained from (9.13) and (9.14) by substituting the expressions (9.5) and (9.6) for r^{g} and V^{g}, respectively:

$$\varphi_{\mathrm{g}}^{\mathrm{LW}}(r,t) = \left| \frac{\rho V}{4\pi\epsilon_0 r} \frac{1}{(1 - \boldsymbol{v}\cdot\boldsymbol{n}/c)} \left(1 - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right) \right|_{\mathrm{ret}} , \tag{9.15}$$

$$\boldsymbol{A}_{\mathrm{g}}^{\mathrm{LW}}(r,t) = \left| \frac{\rho V}{4\pi\epsilon_0 c^2 r} \frac{\boldsymbol{v}}{(1 - \boldsymbol{v}\cdot\boldsymbol{n}/c)} \left(1 - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right) \right|_{\mathrm{ret}} , \tag{9.16}$$

where, as before, the subscript 'ret' indicates that the potentials are evaluated at the retarded time.

We can carry out the same calculations for the anisotropic volume element and the potentials of a charge in an accelerating reference frame N^{a}. For the anisotropic volume element, we can derive the expression

$$V^{\mathrm{a}} = V\left(1 + \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{2c^2}\right) , \tag{9.17}$$

which is in accordance with the equivalence principle, since it can also be obtained from (9.6) by substituting $\boldsymbol{g} = -\boldsymbol{a}$ there.

Then the potential of a charge at rest in N^{a} is easily calculated by taking into account (9.17) and the fact that, due to (8.15) in N^{a}, we also have $r^{\mathrm{a}} = \bar{c}^{\mathrm{a}} t$:

$$\varphi^{\mathrm{a}} = \frac{\rho V}{4\pi\epsilon_0 r}\left(1 + \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\right) . \tag{9.18}$$

As in the case of the anisotropic volume element, the *independent* derivation of (9.18) transforms into (9.7) by substituting $\boldsymbol{a} = -\boldsymbol{g}$, in agreement with the equivalence principle. The same holds for the modified Liénard–Wiechert potentials in N^{a}:

$$\varphi_{\mathrm{a}}^{\mathrm{LW}}(r,t) = \left| \frac{\rho V}{4\pi\epsilon_0 r} \frac{1}{(1 - \boldsymbol{v}\cdot\boldsymbol{n}/c)} \left(1 + \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\right) \right|_{\mathrm{ret}} , \tag{9.19}$$

$$\boldsymbol{A}_{\mathrm{a}}^{\mathrm{LW}}(r,t) = \left| \frac{\rho V}{4\pi\epsilon_0 c^2 r} \frac{\boldsymbol{v}}{(1 - \boldsymbol{v}\cdot\boldsymbol{n}/c)} \left(1 + \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\right) \right|_{\mathrm{ret}} . \tag{9.20}$$

9.3 Calculating the Electric Field of a Charge in a Non-Inertial Reference Frame

The removal of the 1/2 factor in the potential (9.1) derived by Fermi suggests that the same factor will be removed from the electric field (9.2) calculated from that potential. We can now verify this. The electric field of a charge ρV^{g} at rest in N^{g} can be obtained from the scalar potential (9.7):

$$\boldsymbol{E}^{\mathrm{g}} = -\boldsymbol{\nabla}\varphi^{\mathrm{g}} = \frac{\rho V}{4\pi\epsilon_0}\left(\frac{\boldsymbol{n}}{r^2} - \frac{\boldsymbol{g}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} + \frac{\boldsymbol{g}}{c^2 r}\right) , \tag{9.21}$$

where $\boldsymbol{n} = \boldsymbol{r}/r$. The 1/2 factor is not present in (9.21) and the contradiction with the equivalence principle is removed. The advantage of calculating the electric field of a charge at rest in N^{g} directly in N^{g} is that the field is obtained only from the scalar potential (9.7) and no retarded times are involved.

The effect of the average proper velocity (8.14) is that it causes the distortion of the electric field (9.21) in N^{g}. It should be stressed that if the anisotropic propagation of light in N^{g} were not taken into account, an observer there would determine that the field of a charge at rest in N^{g} were the Coulomb field.

The calculation of the electric field of a charge at rest in an accelerating frame N^{a} is also obtained only from the scalar potential (9.18) with no involvement of retarded times:

$$\boldsymbol{E}^{\mathrm{a}} = -\boldsymbol{\nabla}\varphi^{\mathrm{a}} = \frac{\rho V}{4\pi\epsilon_0}\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{a}}{c^2 r}\right) . \tag{9.22}$$

The field (9.22) is derived in N^{a} and coincides with the electric field (9.3) of a charge instantaneously at rest in an inertial reference frame I. (I is the comoving reference frame at a given moment of the time in N^{a}.) The deformation of the electric field (9.3) in I originates from the acceleration of the charge with respect to I. Precisely the same deformation of the field of the charge as determined in N^{a} is caused by the anisotropic velocity of light (8.15) there. This result is another manifestation of the absoluteness of acceleration: an observer in N^{a}, where the charge is at rest, is able to detect the acceleration of N^{a} by determining the shape of the electric field of the charge. That shape should be the same for the observers in N^{a} and I, since they agree that the charge is *accelerating* (with the same acceleration $\boldsymbol{a}$).

The fact that it is the anisotropic propagation of light in N^{a} and N^{g} that causes the distortion of the electric field of a charge at rest in N^{a} and in N^{g} sheds light on another fundamental question: What lies behind the equivalence principle? When the charge is accelerating or supported in a gravitational field, in both cases, its worldline is deviated from its geodesic state; the deformation of the worldlines is the same for $\boldsymbol{a} = -\boldsymbol{g}$. In the immediate vicinity of the worldlines of an accelerating charge and a charge at rest in a gravitational field, everything is the same – the propagation of light in the reference frames N^{a} and N^{g} associated with the equally deformed worldlines of the accelerated charge and the charge supported in a gravitational field (for $\boldsymbol{a} = -\boldsymbol{g}$) is equally anisotropic, and this gives rise to equally distorted fields of the charges and other equal effects, as we will see shortly here and in Chap. 10 as well. One may object that this argument holds only for electromagnetic phenomena. We will see in Chap. 10 that the same argument holds for all interactions. The anisotropic average velocity of propagation of any interaction can be derived in exactly the same way as was done for the electromagnetic interaction.

As accelerated motion is absolute and the shape of the electric field of an accelerating charge is the same for an inertial and a non-inertial observer, it is natural to expect that the field of an inertial charge, which is the Coulomb field, should be the same for an inertial and a non-inertial observer. Let us first check whether this is really the case in N^{a}. Consider an inertial (not accelerating) charge which appears to fall in N^{a} with an apparent acceleration $\boldsymbol{a}^* = -\boldsymbol{a}$, where $\boldsymbol{a}$ is the proper acceleration of N^{a}. Imagine that N^{a} is associated with an accelerating spacecraft in which the charge *appears* to fall. An inertial reference frame I is associated with another spacecraft, which moves by inertia and which is at rest with respect to the charge. *In reality*, the charge is not accelerating (falling) toward the floor of the accelerating spacecraft; it is the floor of that spacecraft which approaches the charge.

The electric field of the falling charge considered instantaneously at rest[4] in N^{a} is obtained from the generalized Liénard–Wiechert potentials (9.19) and (9.20) to within terms proportional to c^{-2}:

$$\begin{aligned}\boldsymbol{E} &= -\boldsymbol{\nabla}\varphi_{\mathrm{a}}^{\mathrm{LW}} - \frac{\partial \boldsymbol{A}_{\mathrm{a}}^{\mathrm{LW}}}{\partial t} \\ &= \frac{e}{4\pi\epsilon_0}\left[\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{a}^*\!\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{a}^*}{c^2 r}\right) + \left(\frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{a}}{c^2 r}\right)\right] .\end{aligned}$$

[4] We consider the instantaneous electric field of the charge in order to separate its length contraction from the distortions caused by (i) its apparent acceleration in N^{a} and (ii) the anisotropic propagation of light there.

Noting that $\boldsymbol{a}^* = -\boldsymbol{a}$, this proves that the instantaneous electric field of the falling charge as described in N^{a} is identical with the field of an inertial charge determined in its rest frame:

$$\boldsymbol{E} = \frac{e}{4\pi\epsilon_0}\frac{\boldsymbol{n}}{r^2} \, .$$

This result shows that both an inertial observer I and a non-inertial observer at rest in N^{a} observe that the instantaneous electric field of the falling charge is the Coulomb field. Comparing the electric field (9.22) of a charge at rest in N^{a} (determined in N^{a}) and its field (9.3) determined in I in which the charge is instantaneously at rest shows that, for both an observer in I and an observer in N^{a}, the field of the charge is equally distorted. There is therefore a unique connection between the shape of the electric field of a charge and its inertial status: if a charge is represented by a geodesic worldline (which means that it moves by inertia) its field is the Coulomb field – both an inertial observer I and a non-inertial observer N^{a} detect the same (Coulomb) field. If the worldline of a charge is not geodesic (meaning that the charge does not move by inertia and resists its acceleration), its electric field is deformed – both I and N^{a} observe the same distorted electric field.

Now consider a charge that is falling in the Earth's gravitational field (i.e., in N^{g}) with an apparent acceleration $\boldsymbol{g}$. For an inertial observer I falling with the charge, its field is not distorted – it is the Coulomb field. The question is whether it will be distorted in N^{g}. As we did in the case of a charge falling in N^{a}, let us calculate the electric field of a charge which is considered instantaneously[5] at rest in N^{g}. It is also obtained from the generalized Liénard–Wiechert potentials (9.15) and (9.16):

$$\begin{aligned}\boldsymbol{E} &= -\boldsymbol{\nabla}\varphi_{\mathrm{g}}^{\mathrm{LW}} - \frac{\partial \boldsymbol{A}_{\mathrm{g}}^{\mathrm{LW}}}{\partial t} \\ &= \frac{\rho V}{4\pi\epsilon_0}\left[\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{g}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{g}}{c^2 r}\right) + \left(-\frac{\boldsymbol{g}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} + \frac{\boldsymbol{g}}{c^2 r}\right)\right] \, ,\end{aligned}$$

which obviously reduces to the Coulomb field

$$\boldsymbol{E} = \frac{\rho V}{4\pi\epsilon_0}\frac{\boldsymbol{n}}{r^2} \, . \tag{9.23}$$

[5] Again, the only reason for considering the instantaneous electric field is to exclude the deformation of the electric field due to length contraction and to examine only the distortions caused by (i) the apparent acceleration of the charge in N^{g} and (ii) the anisotropic propagation of light there.

Therefore the instantaneous electric field of a charge falling in N^{g} is not distorted. This result sheds light on three important questions:

- The connection between the shape of the electric field of a charge and its inertial status. As in the case of a charge described in an accelerating reference frame N^{a}, here too the electric field of a charge falling in N^{g} is the Coulomb field for *both* an inertial observer I falling with the charge and an observer at rest in N^{g}. The electric field (9.21) of a charge at rest in N^{g} and determined there is distorted. The field of a charge at rest in N^{g} determined in an inertial reference frame I falling in N^{g} (and instantaneously at rest in N^{g}) is as distorted as (9.21). We have not calculated it here, but such a calculation is equivalent to calculating the field of an accelerating charge in I [76]; so we have to simply substitute $\boldsymbol{a} = -\boldsymbol{g}$ in (9.3) in order to obtain the field of a charge at rest in N^{g} and determined in I. Therefore the shape of the electric field of a charge does depend on its inertial status and is *absolute* – both an inertial and a non-inertial observer agree on which charge is inertial and which non-inertial by looking at the shape of the fields due to the charge. The field of an inertial charge, whose worldline is geodesic, is the Coulomb field, whereas the field if a non-inertial charge (accelerating or supported in a gravitational field), whose worldline is deformed (not geodesic), is distorted.
- The question as to why, in general relativity, a charge falls in a gravitational field 'by itself' with no force acting on it. As (9.23) shows, the only way for a charge to compensate the anisotropy in the propagation of electromagnetic disturbances and to preserve the Coulomb shape of its electric field is to fall with an acceleration $\boldsymbol{g}$. If the charge is prevented from falling, its electric field distorts and, as we will see in Chap. 10, a self-force caused by the distorted field starts to act on the charge, forcing it to move (fall) in such a way that its Coulomb field is restored. This results in the disappearance of the self-force. In short, the reason why a charge falls in a gravitational field is to prevent its field from being distorted by the anisotropic propagation of light there.[6]

[6] It is tempting to assume that, when a charge is prevented from falling in a gravitational field, its field continues to fall. Then the charge should also fall in order to preserve the Coulomb shape of its field by staying at its center. If this were the case, however, the velocity of any disturbances in the falling electric field would *increase* in a direction parallel to $\boldsymbol{g}$ and decrease in the opposite direction. But as we have seen in Sect. 8.4, light 'falling' in a gravitational field

- The question as to whether or not a charge falling in a gravitational field radiates. The result that the instantaneous electric field of a falling charge is not distorted is a new argument in the debate on whether or not a charge falling in a gravitational field radiates [13, p. 93], [67, 77–82]. It is clear from (9.23) that a falling charge does not radiate since its electric field does not contain the radiation r^{-1} terms. If those terms were present in the field of a falling charge this would constitute a contradiction with the principle of equivalence, since the field of a charge falling in an accelerating frame N^{a} is the Coulomb field and therefore it does not radiate – the charge moves with constant velocity and it is the frame N^{a} that approaches the charge and creates the impression that it is the charge that falls in N^{a}.

9.4 Summary

It has been shown that, by taking into account the anisotropic propagation of light in non-inertial reference frames, the potential and electric field of a charge can be directly calculated in non-inertial reference frames without the need to transform the field from a comoving or local inertial frame.

It has also been shown that the shape of the electric field of a charge depends on its inertial status and is *absolute*, since both an inertial and a non-inertial observer determine that the field of an inertial charge, whose worldline is geodesic, is the Coulomb field, whereas the field of a non-inertial charge, whose worldline is deformed, is distorted.

decelerates. The electric field of a charge in a gravitational field does not fall; it is merely distorted.

10 Inertia as a Manifestation of the Reality of Spacetime

The issues of inertia and gravitation have been the most significant puzzles in physics for centuries. Even now, at the beginning of the twenty-first century, the situation is the same – the nature of inertia remains an unsolved mystery in modern physics and our understanding of gravity can be described in almost the same way, since the modern theory of gravitation, general relativity, has not added much to our *understanding* of the mechanism of the gravitational interaction. General relativity, which provides a surprisingly simple and beautiful no-force explanation for the gravitational interaction of bodies following geodesic paths, remains silent on such important questions as how matter curves spacetime and what is the nature of the very force we identify as gravitational – the force acting upon a body deviated from its geodesic path while being at rest in a gravitational field. The mystery of gravity has been even further highlighted by the fact that the open questions of inertia and gravitation appear to be closely related, since the interpretation of what is called the gravitational force that best reflects the spirit of general relativity suggests that this force is in fact inertial [83]. This naturally explains why "there is no such thing as the force of gravity" in general relativity [84].

Inertia is defined as the resistance a particle offers to its acceleration. This resistance manifests itself as a force – the inertial force $\boldsymbol{F}^{\mathrm{a}}$ – which opposes the external force that accelerates the particle. The inertial force is negatively proportional to the particle's acceleration $\boldsymbol{a}$:

$$\boldsymbol{F}^{\mathrm{a}} = -m^{\mathrm{in}}\boldsymbol{a} ,$$

where the coefficient of proportionality m^{in} is the inertial mass of the particle, which is defined as the *measure* of resistance the particle offers to its acceleration.

The inertial force $\boldsymbol{F}^{\mathrm{a}}$ arises whenever the particle worldtube is deformed, i.e., deviated from its geodesic shape. The gravitational force

$$\boldsymbol{F}^{\mathrm{g}} = m^{\mathrm{g}}\boldsymbol{g} ,$$

where m^{g} is the particle's passive gravitational mass, also arises whenever the worldtube of a particle falling in a gravitational field is deformed. This happens when the particle is prevented from following its geodesic path; for instance, if the particle is on the Earth's surface, it is prevented from falling and its worldtube is deformed (non-geodesic). So the passive gravitational mass is also a measure of the resistance a falling particle offers when prevented from doing so. It appears that the deformation of the worldtube of a non-inertial particle (accelerating or being supported in a gravitational field) gives rise to an inertial force. But before exploring that link further, let us address the still controversial question of the reality of inertial forces.

10.1 Are Inertial Forces Real?

Consider two spacecraft A and B at rest with respect to each other. Imagine there is a ball floating inside A. At a given moment spacecraft A starts to accelerate. We can determine whether any real force is acting on the ball by taking into account how observers in A and B describe what happens to the ball (and of course, we can measure the existence of any force).

An observer in A will see that the ball accelerates while it falls toward the spacecraft's floor and he may assume that a force is acting on the ball and is forcing it to fall. This assumption seems even more probable when the ball hits the floor and starts to exert a force on it (i.e., the ball starts to 'weigh'); it is here that we can measure this force. The A-observer might say that this force is the same force which caused the ball to fall. Then it appears that the A-observer can interpret the nature of that force in two ways – either the force is fictitious (as some claim) or real (as others hold). Both are wrong, due to the incorrect assumption that the *same* force with which the ball acts on the floor is also causing the ball to fall.

Once the A-observer takes into account the fact that acceleration is absolute and that it was spacecraft A that changed its state of motion, the correct explanation becomes evident – no force acts on the 'falling' ball, but at the moment spacecraft A's floor touches the ball and starts to accelerate it, a *real* inertial force arises in the ball, with which it resists its acceleration. That this is really so is best realized when we look at what an observer in B sees. For the B-observer nothing happens to the ball when A starts accelerating since it continues to float in A and does not change its state of motion. It is A's floor that *in reality* approaches the ball. However, things change when the

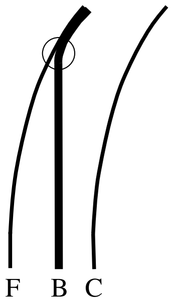


Fig. 10.1. An accelerating spacecraft A is represented by the worldtubes of its floor (F) and its ceiling (C). A ball inside A is depicted by its worldtube B. The collision of the ball and the spacecraft floor is assumed to be plastic

B-observer sees that A's floor reaches the ball. The floor starts to accelerate the ball and the ball resists the change of its state of motion. So a real force appears in the ball which acts back on the floor. It is clear from Fig. 10.1 that the force arises only when the worldtube of the ball is deformed by the worldtube F of spacecraft A's floor.

If the A-observer wants to *describe* the 'fall' of the ball in terms of a force, then he can formally introduce a fictitious force which 'becomes' real at the moment the ball hits the floor. However, the debate over the reality of the inertial forces demonstrates that such a formal approach more often results in confusion.

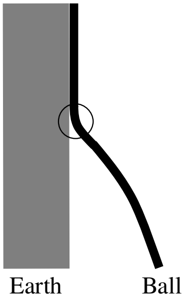


Fig. 10.2. The worldtube of a ball falling toward the Earth's surface is deformed by the worldtube of the Earth. The worldtube of the ball in the upper part of the figure is straight but, as the spacetime in the vicinity of the Earth's worldtube is curved, a straight worldtube is not geodesic and is therefore deformed. The worldtube of the ball in the bottom part of the figure is geodesic and hence is not deformed. The collision of the ball and the Earth's surface is also assumed to be plastic

The same situation occurs with a ball falling toward the surface of the Earth. The ball's worldtube is geodesic, which means that no force acts on the falling ball and it moves by inertia. But a real inertial[1] force appears when the ball reaches the Earth's surface. As in the case of a ball in an accelerating spacecraft, the inertial force arises when the ball's worldtube is deformed. In this case the deformation is caused by the huge worldtube of the Earth, as shown in Fig. 10.2.

The correct description of the behaviour of a ball in the non-inertial reference frame of the spacecraft or the Earth shows that no force is acting on the falling ball and a force arises in the ball when it is deviated from its geodesic state.

10.2 Inertial Forces Originate from a Four-Dimensional Stress Arising in the Deformed Worldtubes of Non-Inertial Bodies

We have seen in Figs. 10.1 and 10.2 that in both cases the inertial force in the ball arises whenever its worldtube is deformed. This observation naturally leads to the question of the *reality* of the ball's worldtube. If it is a real four-dimensional object, then the inertial force turns out to be a restoring force trying to restore the geodesic shape of the deformed worldtube of the ball. This restoring force originates from a four-dimensional stress which arises in the deformed worldtube of the ball.

We see in Fig. 10.1 that the worldtube of the ball is most deformed in the area marked by the circle; that area corresponds to the moment when the ball hits the floor of spacecraft A. (In fact, it is the floor that hits the ball.) Due to the greater deformation of the worldtube in the encircled area, the restoring force there is also greater than the restoring force in the ball's worldtube after the collision between the ball and the floor. And this corresponds exactly to what happens in reality – the inertial force with which the ball acts back on the floor when it is hit by the floor is greater than the inertial force with which the ball resists its acceleration after that moment. The reason is that, due to the relative speed between the ball and the floor, their worldtubes form an angle, as shown in Fig. 10.1, and the ball's worldtube deforms more in the encircled area in order to adjust to the floor's worldtube; after that, the restoring force in the ball's worldtube is smaller, since

[1] The force is inertial because the ball is prevented from moving non-resistantly, i.e., by inertia.

the worldtube of the ball only adjusts to the curvature of the floor's worldtube and as a result its deformation is smaller.

In the case of a ball falling in the Earth's gravitational field (Fig. 10.2) the deformation of the ball's worldtube is again greater in the encircled area which corresponds to the moment when the ball hits the Earth's surface. In that area the restoring force is greater than the restoring force resisting the deformation of the ball's worldtube after the event of the collision between the ball and the Earth's surface. And, again, this is what happens in reality – the ball hits the Earth's surface with a greater force than the force (its weight) it exerts on the surface after the moment of the collision.

The restoring forces which arise in a deformed elastic body that is subjected to deformation can be described in terms of a stress tensor. In the case of the deformed four-dimensional worldtube of a non-inertial particle, the four-dimensional stress tensor is

$$\sigma^{\alpha}_{\beta} = \frac{F^{\alpha}}{V^{\beta}} \, .$$

This represents a force $\boldsymbol{F}$ with components F^{α} in a direction opposite to the direction α in which the external force acts. The external and restoring forces are applied to a three-dimensional volume V^{β} whose normal is in the β direction. As in the case of the deformation of a three-dimensional body, the deformation of a worldtube is described by a strain tensor u^{α}_{β}. The stress tensor is proportional to the strain tensor:

$$\sigma^{\alpha}_{\beta} = \varepsilon u^{\alpha}_{\beta} \, , \tag{10.1}$$

where the coefficient of proportionality ε depends on the elastic properties of the body's worldtube.

Consider the worldtube of a particle of diameter d whose acceleration is in the x^1 direction, which means that the restoring force acts in the opposite direction (Fig. 10.3). The stress tensor is then

$$\sigma^{1}_{0} = \frac{F^{1}}{V^{0}} \, , \tag{10.2}$$

where V^0 is the ordinary volume whose normal is along the time direction. The strain tensor describes the deformation of the particle worldtube:

$$u^{1}_{0} = \frac{\Delta x^{1}}{\Delta x^{0}} = \frac{\Delta x^{1}}{c\Delta t} \, .$$

To determine u^1_0, assume that, at the moment the particle starts to accelerate (the beginning of the deformation of its worldtube), a light

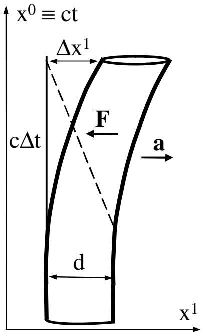


Fig. 10.3. A restoring force $\boldsymbol{F}$ arises in the deformed worldtube of an accelerating particle

signal is emitted from a volume element located at the front end of the particle with respect to its acceleration. After time $\Delta t = d/c$, the light signal arrives at the place where a volume element at the rear end of the particle would be if the particle were not accelerating. Due to the particle's acceleration, the rear end volume element is displaced from its *equilibrium* position by

$$\Delta x^1 = \frac{1}{2} a \Delta t^2 \ ,$$

and a restoring force opposes that displacement. Now we can obtain an explicit expression for the strain tensor:

$$u_0^1 = \frac{\Delta x^1}{c\Delta t} = \frac{1/2(a\Delta t^2)}{c\Delta t} = \frac{d}{2c^2} a \ . \qquad (10.3)$$

Taking into account (10.1), (10.2), and (10.3), the restoring force, which tries to bring all volume elements of the particle back to their equilibrium positions, is

$$F^1 = \frac{\varepsilon V^0 d}{2c^2} a \ .$$

As the restoring force $\boldsymbol{F} = -F^1 \boldsymbol{i}$ and $\boldsymbol{a} = a\boldsymbol{i}$, where $\boldsymbol{i}$ is a unit vector in the x^1 direction, we can write the restoring force in vector form:

$$\boldsymbol{F} = -m^{\mathrm{in}} \boldsymbol{a} \ , \qquad (10.4)$$

where

$$m^{\mathrm{in}} = \frac{\varepsilon V^0 d}{2c^2}$$

can be interpreted as the inertial mass of the accelerating particle, since it is a *measure of the resistance* the particle worldtube offers to its deformation. In terms of the ordinary three-dimensional language, m^{in} is a measure of the resistance the particle offers to its acceleration.

The restoring force (10.4) has exactly the form of the inertial force acting on the accelerating particle. This is an encouraging result, which may explain why inertia has remained a mystery for so long. It has not been realized that the issue of the dimensionality of the world might be closely related to a number of open questions in physics. As in the case of the resolution of another mystery – that there is no absolute motion because the world is four-dimensional[2] – inertia appears to be another manifestation of the four-dimensionality of the world, since it originates from a four-dimensional stress in the deformed worldtube of a non-inertial particle.

What is not so encouraging, however, is that it seems we cannot move further than the derivation of (10.4). In order to go beyond the rather phenomenological treatment of inertia in terms of the four-dimensional stress tensor, we have to examine the interactions that give rise to four-dimensional stress in the deformed worldtube of a non-inertial particle. And here the analogy with a deformed three-dimensional rod helps again. What gives rise to the restoring force and the three-dimensional stress in the rod are *static* electric forces which try to bring all atoms of the rod back to the equilibrium positions which they left when the rod was deformed. So the mechanism responsible for the resistance a three-dimensional rod offers to its deformation is the *displacement* of the rod's atoms from their normal (equilibrium) positions, which the atoms occupy when the rod is not deformed. That displacement disturbs the balance of all electric forces that hold the atoms together in the rod and gives rise to restoring electric forces which try to restore the balance by bringing the atoms back to their equilibrium positions.

This *same* mechanism, which is electromagnetic in origin, appears to give rise to the resistance an accelerating body offers to its acceleration as well. As its atoms are constantly being deviated from their equilibrium positions (as a result of the body's acceleration), one can think of the restoring force in the deformed worldtube of the accelerating body as coming from the static electric forces in the body, which resist the deviation of the body's atoms from the locations they occupy when the body is not accelerating. In other words, the worldlines

[2] If the world were three-dimensional, there should be absolute motion, as we have seen in Chap. 3.

of the atoms comprising the deformed worldtube of the accelerated body are deviated from the shapes they have when the worldtube is not deformed and the electric forces that try to bring the atoms back to their equilibrium positions are attempting to restore the shape of the atoms' worldlines.

At first sight it may appear that inertia originates entirely from the *same* static electric forces that hold the atoms of a solid body together. As the atoms of the body are held at their equilibrium positions by electric forces, when the atoms are deviated from these positions, the same electric forces try to bring the atoms back and this gives rise to a restoring force. However, it turns out that this restoring force cannot be identified with the *entire* inertial force which resists the body's acceleration. The reason is that the strength of the bonds between the atoms in the body are orders of magnitude smaller than the strength needed to account for the entire inertial force. It should be stressed, however, that the restoring static electric forces that try to restore the atoms of an accelerating body to their equilibrium positions do exist, but constitute only a tiny fraction of the inertial force which resists the acceleration of the body.

There is another obvious reason why the electric forces which arise when the atoms of an accelerating body are displaced from their unaccelerated locations cannot be identified with the entire inertial force – these forces do not explain the origin of inertia of *free* atoms, protons, electrons, etc. Those forces cannot account for the inertia of elementary neutral particles either. However, the basic element of the mechanism of the resistance a deformed three-dimensional rod offers whenever it is deformed – that there are restoring forces which try to bring the constituent particles of the rod back to their equilibrium positions – cannot be ignored. The reason is that this restoring mechanism is always present in *any* elementary particle which accelerates and which possesses *any* type of charge – electric, strong, weak, or gravitational.

To show why and how that restoring mechanism is responsible for the inertia of free elementary particles, let us start with electromagnetic interactions by examining the inertia of the electron. The problem is that we do not know what the electron is. That is why we will discuss the two simplest models – the classical model of the electron, according to which the electron is a small spherical shell of negative electric charge, and point-like charges in the Standard Model.

Consider first the classical electron. When the electron is at rest in an inertial reference frame or moves with constant velocity, the mutual repulsion of the elements of the charged spherical shell cancel out exactly and no net force acts on the electron. When the classical electron accelerates, however, all elements of the spherical shell are displaced from their non-accelerated positions and the balance in the mutual repulsion is disturbed. As a result a restoring force arises, which tries to bring all elements of the electron back to their equilibrium positions. Put another way, the accelerating charged spherical shell is *displaced* from the center of its own electric field and sees the field *distorted*. The shell cannot be constantly at the center of the field since it cannot update its field faster than the speed of light. The interaction of the charged shell with its own distorted field produces a restoring or self-force that tries to restore the field to its normal (non-accelerated) shape. Therefore, by resisting the deformation of the electron field, the self-force resists the acceleration of the electron.

The same argument holds for a point-like electron as well. An accelerating point-like electron is displaced from the center of its electric field and as a result sees its field distorted. The interaction of the electron with its own distorted field gives rise to a self-force that resists the deformation of the field. This means that the self-force resists the acceleration of the electron.

As we will see below, the self-force for both the classical and a point-like electron has the form of the inertial force – it is proportional to the acceleration and the coefficient of proportionality is the mass corresponding to the energy of the electric field of the electron through the equation $E = mc^2$. In the case of electromagnetic interactions, that mass is electromagnetic in origin.

If the elementary particles which participate in other interactions are regarded as point-like, it is evident that the *same* mechanism outlined above gives rise to contributions from the other interactions to the four-dimensional stress and the restoring force arising in the deformed worldtube of a non-inertial body. Therefore the resistance a body offers to its acceleration (i.e., its inertia) can be regarded as caused by electromagnetic, weak, and strong interactions in the framework of the Standard Model (which do not include the gravitational interaction). As the mass is a measure of that resistance, it should also be regarded as electromagnetic, weak, and strong in origin.

Before starting our examination of the classical electron, let us address a question which naturally follows from the discussed mechanism of inertia and mass. An electron, for example, is not only displaced from its equilibrium position at the center of its own field when it accelerates. If the electron moves with constant velocity relative to an inertial observer I, it will be constantly deviated from the center of its electric field as seen by I. According to the relativity principle, the electron is always at the center of its field in its rest frame and therefore it is subjected to no net force. As the electron is not accelerating it is not subjected to any force according to I either. But the fact that the electron is displaced from its equilibrium position as determined by I should have some physical meaning. If the electron starts to accelerate, the inertial observer I will find that the displacement from the center of its field will be greater than if determined in the comoving inertial frame. This will be interpreted to mean that the electron offers greater resistance to its acceleration as determined by I. Therefore, I will find that the mass of the electron, being the measure of that resistance, will be *greater*. So the relativistic increase in the mass also appears to follow from the mechanism that is responsible for the inertia and mass of the electron. And indeed, when the momentum of the electric field of a charge moving with constant speed relative to I is calculated, the mass that corresponds to the energy of the charge's field increases with the increase in its speed [74, p. 28-3].

10.3 Electromagnetic Mass and Inertia of the Classical Electron

In 1881 Thomson [85] first realized that a charged particle was more resistant to being accelerated than an otherwise identical neutral particle. He thus conjectured that inertia can be reduced to electromagnetism. Due mainly to the work of Heaviside [86], Searle [87], Abraham [88], Lorentz [89, 90], Poincaré [91, 92], Fermi [71, 93–95], Mandel [96], Wilson [97], Pryce [98], Kwal [99], and Rohrlich [77, 100], this conjecture was developed in the framework of the classical electron theory into what is now known as the classical electromagnetic mass theory of the electron. In this theory inertia is regarded as a local[3] phenomenon originating from the interaction of the electron with its own electromagnetic field.[4]

The electromagnetic mass theory of inertia is still the only theory that predicts the experimental fact that at least part of the inertia and

[3] By contrast, around 1883 Mach [104] argued that inertia was caused by all the matter in the Universe, thus assuming that the local property of inertia had a non-local cause. According to general relativity, however, the only role of all the matter of the Universe is to determine the curvature of spacetime and therefore which worldlines are geodesic.

[4] For the historical development of the classical electromagnetic mass theory, see [77, 101–103]. See also Appendix A.

inertial mass of every charged particle is electromagnetic in origin. As Feynman put it [74]: "There is definite experimental evidence of the existence of electromagnetic inertia – there is evidence that some of the mass of charged particles is electromagnetic in origin." And despite the fact that, at the beginning of the twentieth century, many physicists recognized "the tremendous importance which the concept of electromagnetic mass possesses for all of physics" since "it is the basis of the electromagnetic theory of matter" [94], it has been inexplicably abandoned after the advent of relativity and quantum mechanics. And that happened even though the classical electron theory predicted that the electromagnetic mass increases with increasing velocity *before* the theory of relativity, yielding the correct velocity dependence, and that the relation[5] between energy and mass is $E = mc^2$ [74, pp. 28-3, 28-4]. Today [105, p. 213] "the state of the classical electron theory reminds one of a house under construction that was abandoned by its workmen upon receiving news of an approaching plague. The plague in this case, of course, was quantum theory. As a result, classical electron theory stands with many interesting unsolved or partially solved problems."

Often the first reaction to any study of the classical model of the electron questions why it should be studied at all since it is clear that this model is wrong. The classical electron radius that gives the correct electron mass is $\sim 10^{-15}$ m, whereas experiments probing the scattering properties of the electron find that its size is smaller than 10^{-18} m [106]. Unfortunately (or fortunately), it is not that simple. An analysis of why the electron does not appear to be so small has been carried out by MacGregor [107]. However, what immediately shows that the scattering experiments do not tell the whole story is the fact that they are relevant only to the particle aspect of the electron.

Despite all the studies specifically devoted to the nature of the electron (see for instance [108, 109]), no one knows what an electron looks like before being detected and some even deny the very correctness of such a question. One thing, however, is completely clear – the experimental upper limit of the size of the electron ($< 10^{-18}$ m) cannot be interpreted to mean that the electron is a particle (localized in a region whose size is smaller than 10^{-18} m) without contradicting both quantum mechanics and the existing experimental evidence. (Recall our discussion of the dipole moment of a hydrogen atom in Chap. 6.) Therefore, the scattering experiments tell very little about what the electron actually is and further studies will be needed in order to un-

[5] In fact, the relationship obtained between mass and energy contained a factor of 4/3, which was later accounted for; see [74, p. 28-4].

derstand their meaning. For this reason those experiments are not an argument against any study of the classical model of the electron.

As one of the most difficult problems of the classical electron is its stability (what holds its charge together), one may conclude that the basic assumption in the classical model of the electron – that there is *interaction* between the elements of its charge through their distorted fields – may be wrong. The very existence of the radiation reaction force, however, seems to imply that there is indeed interaction (repulsion) between the different 'parts' of the electron charge [72, p. 439]: "The radiation reaction is due to the force of a charge on itself – or, more elaborately, the net force exerted by the fields generated by different parts of the charge distribution acting on one another." In the case of a *single* radiating electron the presence of a radiation reaction force appears to suggest that there is interaction between different 'parts' of the electron.

Here we shall not follow the standard approach to calculating the self-force [72, 73, 76, 110] which describes an electron's accelerated motion in an inertial frame I. Instead, all calculations will be carried out in the non-inertial reference frame N^{a} in which the accelerating electron is at rest. The reason for this is that the calculation of the electric field and the self-force of an accelerating electron in the accelerating frame N^{a} (not in I) is essential for the correct application of the principle of equivalence, since it relates those quantities of an electron in a non-inertial (accelerating) frame N^{a} and in a non-inertial frame N^{g} supported in a gravitational field. An advantage of calculating the electron's electric field in the non-inertial frame in which the electron is at rest is that it is obtained only from the scalar potential and the calculations do not involve retarded times, as we have seen in Chap. 9.

Taking into account the anisotropic volume element (9.17),

$$\mathrm{d}V^{\mathrm{a}} = \mathrm{d}V\left(1 + \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{2c^2}\right) , \tag{10.5}$$

we have for the scalar potential of a charged volume element $-\rho \mathrm{d}V^{\mathrm{a}}$ of the electron at rest in N^{a},

$$\mathrm{d}\varphi^{\mathrm{a}} = -\frac{\rho}{4\pi\epsilon_0 r}\left(1 + \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\right)\mathrm{d}V . \tag{10.6}$$

Making use only of the scalar potential (10.6), we obtain the electric field of the charged volume element $-\rho \mathrm{d}V^{\mathrm{a}}$ at rest in N^{a} as

$$\mathrm{d}\boldsymbol{E}^{\mathrm{a}} = -\boldsymbol{\nabla}\mathrm{d}\varphi^{\mathrm{a}} = -\frac{1}{4\pi\epsilon_0}\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{a}}{c^2 r}\right)\rho\mathrm{d}V \ .$$

The electric field of the electron is then

$$\boldsymbol{E}^{\mathrm{a}} = -\frac{1}{4\pi\epsilon_0}\int\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{a}}{c^2 r}\right)\rho\mathrm{d}V \ . \qquad (10.7)$$

The self-force which the field of the electron exerts upon an element $-\rho\mathrm{d}V_1^{\mathrm{a}}$ of its own charge is

$$\mathrm{d}\boldsymbol{F}_{\mathrm{self}}^{\mathrm{a}} = -\rho\mathrm{d}V_1^{\mathrm{a}}\boldsymbol{E}^{\mathrm{a}} = \frac{1}{4\pi\epsilon_0}\int\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{a}}{c^2 r}\right)\rho^2\mathrm{d}V\mathrm{d}V_1^{\mathrm{a}} \ . \qquad (10.8)$$

The resultant self-force acting on the electron as a whole is

$$\boldsymbol{F}_{\mathrm{self}}^{\mathrm{a}} = \frac{1}{4\pi\epsilon_0}\int\int\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{a}}{c^2 r}\right)\rho^2\mathrm{d}V\mathrm{d}V_1^{\mathrm{a}} \ , \qquad (10.9)$$

which becomes, after taking into account the anisotropic volume element (10.5),

$$\boldsymbol{F}_{\mathrm{self}}^{\mathrm{a}} = \frac{1}{4\pi\epsilon_0}\int\int\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{\boldsymbol{a}}{c^2 r}\right)\left(1 + \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{2c^2}\right)\rho^2\mathrm{d}V\mathrm{d}V_1 \ .$$

Assuming a spherically symmetric distribution of the electron charge [90] and following the standard procedure for calculating the self-force [110], we get (see Appendix B)

$$\boldsymbol{F}_{\mathrm{self}}^{\mathrm{a}} = -\frac{U}{c^2}\boldsymbol{a} \ , \qquad (10.10)$$

where

$$U = \frac{1}{8\pi\epsilon_0}\int\int\frac{\rho^2}{r}\mathrm{d}V\mathrm{d}V_1$$

is the energy of the electron's electric field. As U/c^2 is the mass that corresponds to that energy, we can write (10.10) in the form

$$\boldsymbol{F}_{\mathrm{self}}^{\mathrm{a}} = -m^{\mathrm{a}}\boldsymbol{a} \ , \qquad (10.11)$$

where $m^{\mathrm{a}} = U/c^2$ is identified as the electron inertial electromagnetic mass. The famous factor of 4/3 in the electromagnetic mass of the electron does not appear in (10.11). The reason is that, in (10.8) and (10.9), we have identified and used the correct volume element $\mathrm{d}V_1^{\mathrm{a}} = \left(1 + \boldsymbol{a}\cdot\boldsymbol{r}/2c^2\right)\mathrm{d}V_1$ originating from the anisotropic velocity of light in N^{a}; not taking it into account results in the appearance of the 4/3 factor.

The self-force $\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}}$ to which an electron is subjected due to its own distorted field is directed opposite to $\boldsymbol{a}$ and therefore resists the acceleration of the electron. As can be seen from (10.10), this force is purely electromagnetic in origin and therefore both the resistance the classical electron offers to being accelerated (i.e., its inertia) and its inertial mass (which is the measure of that resistance) are purely electromagnetic in origin as well.

The self-force (10.11) is traditionally called the inertial force. According to Newton's third law, the external force $\boldsymbol{F}$ that accelerates the electron and the self-force $\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}}$ which resists $\boldsymbol{F}$ have equal magnitudes and opposite directions: $\boldsymbol{F} = -\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}}$. Therefore we can write $\boldsymbol{F} = m^{\mathrm{a}}\boldsymbol{a}$ which means that Newton's second law can be *derived* on the basis of Maxwell's electrodynamics applied to the classical model of the electron and Newton's third law.

In Chap. 9 we calculated the electric field of a charge which appears to fall in N^{a} with an apparent acceleration $\boldsymbol{a}^{*} = -\boldsymbol{a}$ (where $\boldsymbol{a}$ is the proper acceleration of N^{a}). We found that, at every moment, it is the Coulomb field. This result shows that, for a non-inertial observer at rest in N^{a}, the instantaneous electric field of a falling electron is not distorted, and this in turn shows that *no* self-force acts on the electron. Therefore, our calculation of the self-force, demonstrating that the classical electron is subjected to a self-force only when its field is distorted, confirms the result of Chap. 9 that the shape of the electric field of an inertial electron is absolute – both an inertial observer I falling with the electron and a non-inertial observer at rest in N^{a} detect a Coulomb field of the falling electron. In general:

- a Coulomb field is associated with an inertial electron (represented by a geodesic worldline) by both an inertial observer I (moving with the electron) and a non-inertial observer N^{a},
- for both I and N^{a}, the electric field of a non-inertial electron (whose worldline is not geodesic) is equally distorted.

As we expected the fact that the state of (inertial or accelerated) motion of a charge is absolute implies that the shape of the electric field of an (inertial or accelerated) charge is also absolute (the same for an inertial and a non-inertial observer).

Similarly to the case of calculating the electric potential in N^{a}, the average anisotropic velocity of light (8.14) in N^{g} also leads to anisotropic r^{g} and $\mathrm{d}V^{\mathrm{g}}$ in N^{g}, as we saw in Chap. 9:

$$(r^{\mathrm{g}})^{-1} \approx r^{-1}\left(1 - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{2c^2}\right)$$

and

$$\mathrm{d}V^{\mathrm{g}} = \mathrm{d}V\left(1 - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{2c^2}\right) . \tag{10.12}$$

As a result the scalar potential of a charged volume element $-\rho\mathrm{d}V^{\mathrm{g}}$ of the electron in N^{g} is

$$\mathrm{d}\varphi^{\mathrm{g}} = -\frac{\rho}{4\pi\epsilon_0 r}\left(1 - \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right)\mathrm{d}V . \tag{10.13}$$

Recall that just by taking into account the anisotropic volume element $\mathrm{d}V^{\mathrm{g}}$, we can obtain the correct potential (10.13) of a charge supported in a gravitational field.

The calculation of the electric field of a charged volume element $-\rho\mathrm{d}V^{\mathrm{g}}$ in N^{g} is again carried out using only the scalar potential (10.13):

$$\mathrm{d}\boldsymbol{E}^{\mathrm{g}} = -\boldsymbol{\nabla}\mathrm{d}\varphi^{\mathrm{g}} = -\frac{1}{4\pi\epsilon_0}\left(\frac{\boldsymbol{n}}{r^2} - \frac{\boldsymbol{g}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} + \frac{\boldsymbol{g}}{c^2 r}\right)\rho\mathrm{d}V ,$$

and the field of the electron is then

$$\boldsymbol{E}^{\mathrm{g}} = -\frac{1}{4\pi\epsilon_0}\int\left(\frac{\boldsymbol{n}}{r^2} - \frac{\boldsymbol{g}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} + \frac{\boldsymbol{g}}{c^2 r}\right)\rho\mathrm{d}V . \tag{10.14}$$

A comparison of the electric field of an electron supported in the Earth's gravitational field (10.14), determined in N^{g}, with the electric field of an accelerated electron (10.7), determined in the frame N^{a}, indicates that the electric fields of an electron at rest on the Earth's surface and an electron at rest in the frame N^{a} which moves with an acceleration $\boldsymbol{a} = -\boldsymbol{g}$ are equally distorted in accordance with the principle of equivalence. Substituting $\boldsymbol{a} = -\boldsymbol{g}$ into the electric potential (10.6) also transforms it into (10.13), as required by the equivalence principle.

The self-force with which the electron field interacts with an element $-\rho\mathrm{d}V_1^{\mathrm{g}}$ of the electron charge is therefore

$$\mathrm{d}\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}} = -\rho\mathrm{d}V_1^{\mathrm{g}}\boldsymbol{E}^{\mathrm{g}} = \frac{1}{4\pi\epsilon_0}\int\left(\frac{\boldsymbol{n}}{r^2} - \frac{\boldsymbol{g}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} + \frac{\boldsymbol{g}}{c^2 r}\right)\rho^2\mathrm{d}V\mathrm{d}V_1^{\mathrm{g}} , \tag{10.15}$$

and the resulting self-force with which the electron acts upon itself is

$$\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}} = \frac{1}{4\pi\epsilon_0}\int\int\left(\frac{\boldsymbol{n}}{r^2} - \frac{\boldsymbol{g}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} + \frac{\boldsymbol{g}}{c^2 r}\right)\rho^2\mathrm{d}V\mathrm{d}V_1^{\mathrm{g}} . \tag{10.16}$$

After taking into account the explicit form (10.12) of $\mathrm{d}V_1^{\mathrm{g}}$, assuming a spherically symmetric distribution of the electron charge, and calculating the self-force as we have done in the case of an electron at rest in N^{a}, we get

$$\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}} = \frac{U}{c^2}\boldsymbol{g} \ , \tag{10.17}$$

where

$$U = \frac{1}{8\pi\epsilon_0}\int\int\frac{\rho^2}{r}\mathrm{d}V\mathrm{d}V_1$$

is the energy of the electron's field. As U/c^2 is the mass associated with the field energy of the electron, i.e., its electromagnetic mass, (10.17) takes the form

$$\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}} = m^{\mathrm{g}}\boldsymbol{g} \ , \tag{10.18}$$

where $m^{\mathrm{g}} = U/c^2$ is interpreted here as the electron passive gravitational mass. As in the case of the self-force acting on an accelerating electron described in N^{a}, the 4/3 factor in the electromagnetic mass does not appear in (10.18) for the same reason: the correct volume element (10.12) was used in (10.16). Therefore the anisotropic volume element $\mathrm{d}V^{\mathrm{g}}$ *simultaneously* resolves two different problems – it removes both the 1/2 factor in the potential (9.1) derived by Fermi, as we have seen in Chap. 9, and the 4/3 factor in the self-force.

The self-force $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}}$ which acts upon an electron on account of its own distorted field is directed parallel to $\boldsymbol{g}$ and resists the deformation of its electric field caused by the fact that the electron at rest on the Earth's surface is prevented from falling. This force is traditionally called the gravitational force. As we have seen, $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}}$ arises only when an electron is prevented from falling, i.e., only when it is deviated from its geodesic path. Only in this case does the electron field deform, giving rise to the self-force $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}}$. Thus $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}}$ resists the deformation of the field of the electron, which means that it resists its being prevented from following a geodesic path. As a Coulomb field is associated with a non-resistantly moving electron (represented by a geodesic worldline) it follows that $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}}$ is, in fact, an *inertial* force, since it resists the deviation of an electron from its geodesic path, that is, $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}}$ resists the deviation of the electron from its motion by inertia.

Consequently, the nature of the force acting upon the classical electron at rest in a gravitational field is inertial and is purely electromagnetic in origin as seen from (10.18), which means that the electron passive gravitational mass m^{g} in $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}} = m^{\mathrm{g}}\boldsymbol{g}$ is also purely electromagnetic in origin. It is immediately clear from this why the inertial and the passive gravitational masses of the classical electron are equal.

As the nature of the self-force $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}}$ is inertial, it follows that what is traditionally called passive gravitational mass is, in fact, inertial mass. This becomes evident from the fact that the two masses are the measure of resistance an electron offers when deviated from its geodesic path. In flat spacetime the force of resistance is $\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = -m^{\mathrm{a}}\boldsymbol{a}$, whereas in curved spacetime this same force that resists the deviation of the electron from its geodesic path is $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}} = m^{\mathrm{g}}\boldsymbol{g}$, where m^{a} and m^{g} are the measures of resistance (inertia) in both cases. The two resistance forces are equal $F^{\mathrm{a}}_{\mathrm{self}} = F^{\mathrm{g}}_{\mathrm{self}}$ for $\boldsymbol{a} = \boldsymbol{g}$ as can be seen from (10.10) and (10.17), and therefore $m^{\mathrm{a}} = m^{\mathrm{g}}$. This equivalence also follows from the fact that m^{a} and m^{g} are the *same* thing – the mass associated with the energy of the electron field.

The result that the force $\boldsymbol{F}^{\mathrm{g}}_{\mathrm{self}}$, acting on an electron when it is deviated from its geodesic path due to its being at rest in a gravitational field, is inertial is valid not only for the classical electron. A non-resistant motion (i.e., motion by inertia) of a body in both special relativity (flat spacetime) and general relativity (curved spacetime) is represented by a geodesic worldline, whereas a body represented by a non-geodesic worldline is subjected to a resistance force which opposes the external force preventing the body from following its geodesic path in spacetime. That is why the resistance force is inertial in origin in both special and general relativity. Note that the conclusion of the non-gravitational nature of the force acting on a body at rest in a gravitational field follows from general relativity itself – as a body supported in a gravitational field is deviated from its geodesic path, which means that it is prevented from moving non-resistantly (by inertia), it is subjected to an inertial force which arises only when the body is prevented from moving in a non-resistant (inertial) manner.

According to general relativity, the worldline of an electron falling in a gravitational field is geodesic. This implies that the electron moves by inertia (without resistance) and its Coulomb field is not distorted for an inertial observer I falling with the electron and should not be distorted for an observer in N^{g} either. In Chap. 9, we saw that this is really the case – at any instant, the field of a falling electron is not distorted in N^{g} which, in the light of the results of the present chapter, means that it is subjected to no self-force; that is, the electron offers no resistance to its accelerated motion in N^{g}. This confirms the result in Chap. 9 as to why the electron is falling in a gravitational field *by itself*, while no external force is causing its acceleration – the only way for the electron to compensate the anisotropy in the propagation of light and to preserve the Coulomb shape of its electric field is to fall with an acceleration $\boldsymbol{g}$. If the electron is prevented from falling, its electric field distorts and the self-force (10.18) appears and tries to force the electron to move (fall) in such a way that its Coulomb field is restored; when the distortion of the electron field is eliminated, the self-force disappears.

The result that the classical electron falls in a gravitational field with an acceleration $\boldsymbol{g}$ *by itself* in order to prevent its field from getting distorted is valid for any electric charge, as well as for the weak and strong charges. Consider point-like electric, weak, and strong charges which are supported in a gravitational field. Due to the anisotropic propagation of the three interactions in the gravitational field, the fields of their charges are distorted; that is, the charges are no longer at their equilibrium positions at the center of their fields. As a result, a self-force appears and tries to move the charges to their equilibrium positions. If the charges are left free, that self-force will make them accelerate with an acceleration $\boldsymbol{g}$ because only in this case is the anisotropy in the propagation of the electromagnetic, weak, and strong interactions compensated, so that the charges see their fields undistorted. This mechanism sheds additional light on two facts:

- that in general relativity the motion of a body falling toward a gravitating center is regarded as inertial (non-resistant) and is represented by a geodesic worldline,
- the experimental fact that all objects fall in a gravitational field with the *same* acceleration.

We are now in a position to summarize the relation between the motion of the classical electron and the shape of its field:

- An electron moving with constant velocity in a flat spacetime region where the average velocity of light is isotropic does not resist its uniform motion since uniform motion represented by a straight worldline in flat spacetime ensures that the electron's electric field is the Coulomb field. Stated another way, the only way for an electron to prevent its electric field from distorting in flat spacetime is to move with constant velocity.
- An accelerating electron resists its acceleration in flat spacetime because the accelerated motion distorts the electron's electric field and this results in an electric self-force that opposes the deformation of the electron's field.
- An electron falling toward the Earth's surface does not resist its (flat-spacetime) acceleration since, as we have seen, by falling with an acceleration $\boldsymbol{g}$, the electron compensates the anisotropy in the

propagation of light in the Earth's vicinity and prevents its electric field from getting distorted (the curved-spacetime acceleration of the falling electron is zero).

- An electron at rest on the Earth's surface is subjected to an electric self-force trying to make the electron fall since the average anisotropic velocity of light in the Earth's gravitational field distorts the electron field and this in turn gives rise to the self-force. The nature of that force is inertial (not gravitational) and is electromagnetic in origin.

The mechanism giving rise to the free (non-resistant) fall of the electron in a gravitational field and to the self-force (10.18) that resists its prevention from falling is identical to the mechanism responsible for the self-force (10.10) an accelerated electron in flat spacetime is subjected to and for its free fall as described in the accelerated frame N^{a}. This mechanism implies that the identical anisotropy in the propagation of light in N^{a} and N^{g} gives rise to similar phenomena, whose mathematical expressions transform into one another when the substitution $\boldsymbol{a} = -\boldsymbol{g}$ is used. As we have seen in Sect. 8.4, the identical anisotropy in the propagation of light in N^{a} and N^{g} originates from the fact that the worldlines of bodies at rest in N^{a} and in N^{g} are *equally* deviated from their geodesic shapes. The equivalence principle itself appears to originate from this fact since it gives rise to the anisotropic propagation of light in N^{a} and N^{g}, which in turn leads to *identical* electromagnetic phenomena in N^{a} and N^{g}. In Sect. 8.4, we have seen that the propagation of weak and strong interactions is also anisotropic in N^{a} and N^{g}. This anisotropy also causes identical weak and strong phenomena there.

We have seen that the inertial and passive gravitational masses of the classical electron are entirely electromagnetic in origin. As all three masses – inertial, passive gravitational, and active gravitational – are considered equal,[6] it follows that the classical electron's active gravitational mass is fully electromagnetic in origin as well. And since it is only the charge of the classical electron that represents it (there is

[6] The equivalence of the active and passive gravitational masses can be demonstrated in the following way. Consider a particle on the surface of the Earth. The gravitational force on the particle is given by Newton's second law $F = mg$, where m is its *passive* gravitational mass. The same force can be written in terms of Newton's gravitational law $F = GMm/r^2$, where G is the gravitational constant, r is the Earth's radius, and m and M are the *active* gravitational masses of the particle and Earth, respectively. But this equation can be written in the form $F = m\left(GM/r^2\right) = mg$, where the mass m plays the role of a passive gravitational mass.

no mechanical mass), it follows that the active gravitational mass of the electron is represented by its charge. Therefore it is the electron charge that distorts spacetime and causes the average anisotropic velocity of light in the electron's neighborhood.

As it is the anisotropy in the average proper velocity of light and the electromagnetic mass theory that fully and consistently explain the fall of the classical electron toward the Earth and the self-force acting on an electron at rest on the Earth's surface, it appears natural to expect the gravitational attraction between two electrons to be explained in the same way.

In addition to the electric repulsion of two electrons (e_1 and e_2) in space, they also attract each other through the anisotropy in the average proper velocity of light caused by each of them: e_1 falls toward e_2 in order to compensate the anisotropy caused by e_2 and to prevent its electric field from getting distorted and vice versa. In other words, the charge of the electrons affects the propagation of light around them, which in turn changes the shape of the electron worldlines resulting in their convergence toward each other. Therefore, in the framework of the electromagnetic mass theory, the anisotropy in the propagation of light in the electrons' vicinity is sufficient to explain their mutual (gravitational) attraction in terms of non-resistant motion, which is not caused by a force. In such a way, as we have seen above, the case of the classical electron may provide further insight into the question of why no force is involved in the gravitational attraction of bodies as described by general relativity.

10.4 The Standard Model and Inertia

The study of the classical electromagnetic mass theory makes it possible to ask whether the *same* mechanism that accounts for the inertia and mass of the classical electron – the interaction of the electron charge with its own distorted field – also leads to contributions to inertia and mass from the other interactions in the framework of quantum field theory. As we will see, the same type of acceleration-dependent self-interaction effects are also present in quantum field theory and appear to give rise to contributions from the electromagnetic, weak, and strong interactions to inertia and mass. Since there has been no success in quantizing the gravitational field, let us examine the question of the origin of inertia in the Standard Model [111], which does not consider the effects of gravitational interactions on the behavior of fundamental particles.

All interactions in the Standard Model are realized through the exchange of virtual quanta that constitute the corresponding 'fields' of the interacting particles. For example, in the case of electromagnetic interactions the quantized electromagnetic field of a charge is represented by a cloud of virtual photons which are being constantly emitted and absorbed by the charge. The electric forces of attraction and repulsion between two charges interacting through exchange of virtual photons originate from the *recoils* the charges suffer when the virtual photons are emitted and absorbed. A free (inertial) charge in the Standard Model is subjected to the recoils resulting from the emitted and absorbed virtual photons which constitute its own electric field. Due to spherical symmetry, all recoils caused by both the emitted and absorbed virtual photons cancel out exactly and the charge is not subjected to any self-force. The quantized electromagnetic field of a charge is an analog of an undistorted field, if the self-interaction of the charge and its field produces no self-force acting on the charge. Hence, in terms of the Standard Model, a charge moves non-resistantly (by inertia) if the recoils from the emitted and absorbed virtual photons completely cancel out. In other words, as in the case of the classical electron, a charge whose field is not distorted is represented by a geodesic worldline.

As it is the momentum of a photon that determines the recoil felt by a charge when the photon is emitted or absorbed, the recoils resulting from the virtual photons emitted by a non-inertial charge also cancel out since, as seen by the charge, all photons are emitted with the same frequencies (and energies) and therefore the same momenta. However, the frequencies of the virtual photons coming from different directions before being absorbed by a *non-inertial* charge are direction dependent (general-relativistically blue- or redshifted). The directional dependence of the frequencies of the incoming virtual photons constitutes a distortion of the electric field of a non-inertial charge. The mechanism responsible for the distortion of the field is the same mechanism that follows from the assumption that inertia originates from a four-dimensional stress which arises in the deformed worldtube of a non-inertial charge – the charge is displaced from its equilibrium position at the center of its own field.

It has not been noticed so far that the directional dependence of the frequencies of the virtual photons absorbed by a non-inertial charge disturbs the balance of the recoils to which the charge is subjected. In turn, that imbalance gives rise to a self-force which acts on the non-inertial charge. The self-force resulting from the imbalance in the recoils caused by the virtual photons absorbed by a non-inertial charge is a resistance force since it acts only on non-inertial charges. It arises only when an inertial charge is prevented from following a geodesic worldline; no self-force acts on an inertial charge which follows a geodesic worldline. As in the case of the classical electron, the worldline of a non-inertial charge, whose field (of virtual quanta) is distorted, is not geodesic. As we shall see shortly, in the case of an accelerating charge, the resistance self-force has the form of the inertial force which resists the deviation of the charge from its geodesic path. When a charge is supported in a gravitational field, a self-force, which has the form of what is traditionally called the gravitational force, also resists the deviation of the charge from its geodesic path.

It is clear that non-inertial weak and strong (color) charges will also be subjected to a self-force arising from the imbalance in the recoils caused by the absorbed W and Z bosons, in the case of weak interactions, and the absorbed gluons, in the case of strong interactions. Therefore the acceleration-dependent self-interaction effects to which a non-inertial particle is subjected in the Standard Model are similar to those in the classical electromagnetic theory, and may account for the origin of inertia and mass. As the Standard Model does not account for the masses (and inertia) of fundamental particles, it is believed that a fifth interaction is needed to explain how the masses are generated.[7] The acceleration-dependent self-interaction mechanism outlined above appears to explain the origin of inertia and mass in the framework of the Standard Model without the need for any extra interactions.

The picture which emerges is the following. Consider a body whose constituents are subjected to electromagnetic, weak, and strong interactions. If the recoils from all virtual quanta (photons, W and Z bosons, and gluons) mediating the interactions cancel out precisely, the body is represented by a geodesic worldline; it offers no resistance to its motion and therefore moves by inertia. When the body is accelerating, the balance of the recoils caused by the absorbed virtual quanta is disturbed and this gives rise to a self-force. The worldline of a body whose constituents have distorted electromagnetic, weak, and strong fields is not geodesic. (As in the case of a quantized electromagnetic field, the distortion of the fields manifests itself in the fact that the recoils from the absorbed virtual quanta do not cancel out.) As a result the body resists the deformation of its electromagnetic, weak,

[7] This fifth interaction and the corresponding fifth force – whose simplest version is the Higgs force – is expected by many to have a better fate than the fifth force that was proposed in the late 1980s.

and strong fields and therefore resists its acceleration, which is causing the deformation. The self-force is a resistance force and is composed of three components – electromagnetic, weak, and strong. Therefore both inertia and inertial mass appear to originate from the lack of cancellation of the recoils caused by the absorbed virtual quanta mediating the electromagnetic, weak, and strong interactions.

If the body is at rest in a gravitational field, the frequencies (i.e., the energies) of the virtual quanta being *absorbed* by its constituent particles are general-relativistically shifted (as compared to the frequencies of the virtual quanta that are absorbed by an inertial particle). As a result, the recoils from the virtual photons, W and Z bosons, and gluons which every constituent particle of the body suffers do not cancel out. That imbalance in the recoils gives rise to a self-force which, as we will see, has the form of the gravitational force and is also composed of three components – electromagnetic, weak and strong. This means that the passive gravitational mass, which is the measure of resistance a particle offers when prevented from falling in a gravitational field, also (like the inertial mass) appears to originate from the imbalance in the recoils caused by the absorbed virtual quanta mediating the electromagnetic, weak, and strong interactions. Therefore the same acceleration-dependent self-interaction mechanism provides a natural explanation of the equivalence of inertial and passive gravitational masses, not only in the classical electron theory, but in the Standard Model as well – these masses are not just equivalent; they are the *same* thing, since they have the *same* origin. The anisotropy in the propagation of the virtual quanta is compensated if the body falls with an acceleration $\boldsymbol{g}$. The recoils from all absorbed virtual quanta (virtual photons, W and Z bosons, and gluons) the falling body suffers cancel out exactly and the body moves in a non-resistant way (following a geodesic path). In the framework of the Standard Model, this mechanism offers a *common* explanation of why all bodies fall in a gravitational field with the *same* acceleration and why they do *not* resist their fall.

The picture outlined here suggests that inertia and the entire inertial and passive gravitational mass originate from acceleration-dependent self-interaction effects in the Standard Model – the constituent particles of every non-inertial body are subjected to a self-force which is caused by the imbalance in the recoils from the absorbed virtual quanta. What appears to spoil the picture are the rest masses of the W and Z bosons. As these particles are the carriers of the weak force, the unbalanced recoils from them explains the contribution of the weak interaction to the inertia and mass of every particle undergoing weak interactions in which the W and Z bosons are involved. But what accounts for the inertia and mass of the very carriers of the weak interaction? In fact, the W^+ an W^- bosons do not pose a real problem since they are subjected to the recoils of virtual photons due to their electric charge and therefore their mass might turn out to be entirely electromagnetic in origin. However, the origin of inertia and mass of the neutral Z boson still remains a mystery.

On the one hand, it does follow from the Standard Model, when the general relativistic frequency shift is taken into account, that electromagnetic, weak, and strong interactions all make contributions to the inertia and mass of fundamental particles. On the other hand, the fact that the Z boson involved in mediating the weak interaction possesses rest mass implies that not all mass is composed of electromagnetic, weak, and strong contributions. Obviously, it will be experiment that ultimately determines how much of the mass is due to electromagnetic, weak, and strong interactions, and how much is caused by the Higgs or another unknown mechanism.

One obvious question that has remained unanswered so far is whether or not the gravitational interaction contributes to the inertia and mass of the particles. If we manage to quantize the gravitational field and the existence of gravitons is confirmed, gravitational interaction will make a contribution to inertia and mass as well, and perhaps may even account for the mass of the Z boson.

It should be specifically stressed, however, that the proposed mechanism which gives rise to the self-force acting on a non-inertial particle is *not* hypothetical. It follows directly from the accepted mechanism responsible for the origin of attraction and repulsion forces in the Standard Model when the general-relativistic directional dependence of the frequencies of the *incoming* virtual quanta is taken into account. Therefore the Standard Model does say something important about the origin of inertia and mass.

At the conceptual level it is certain that, in the Standard Model, the recoils from all incoming (blue- or redshifted) virtual quanta absorbed by the electric, weak, and strong non-inertial charges do not cancel out and give rise to an inertial force. However, a conceptual analysis alone is not sufficient for any advancement in physics.

What is not only difficult, but almost hopeless, is how one can calculate the self-force originating from the unbalanced recoils from the virtual quanta that are absorbed by a non-inertial particle. There are too many unknowns. Virtual quanta are off-mass-shell particles, which means that their energy, momentum and mass do not obey the relativistic equation:

$$E^2 = p^2c^2 + (mc^2)^2 \ .$$

Unlike the momentum of a normal photon which is given by $p = E/c$, the momentum of a virtual photon is not equal to E/c. Also unknown are the energy of the field of a charge (electromagnetic, weak, or strong) in terms of the energies of the virtual quanta that constitute the 'field' of the charge, the lifetimes and energies of individual virtual quanta, and the absorption time during which a virtual quantum is absorbed by the charge.

In an attempt to overcome those difficulties we will carry out a semiclassical calculation of the self-force in the case of the electromagnetic interaction in quantum electrodynamics; similar calculations can be done for the other interactions. For this purpose the following assumptions will be made. It appears natural to define the energy of the electric field of a charge in quantum electrodynamics as the sum of the energies of all virtual photons constituting the field at a given moment. Such a definition avoids the issue of the dimension of the charge. In order to eliminate some of the unknowns mentioned above, it appears that an equivalent definition of the charge's field energy is also possible in terms of the total energy of the number of virtual photons absorbed by the charge during some characteristic time δt. (This δt can be regarded as the time over which the charge renews its field.) The lifetimes of all virtual photons can be expressed in terms of δt as $\alpha\delta t$, where α is a real number. The distances travelled by the virtual photons during their lifetimes will then be $\alpha r = \alpha c\delta t$, where $r = c\delta t$, is obviously the distance travelled by a virtual photon during the characteristic time δt. The time during which a virtual photon is absorbed by a charge is assumed to be $\delta\tau$. The momentum of a virtual photon can be considered to be $p^{\mathrm{a}} = \beta\,(E^{\mathrm{a}}/c)$, where β is also a real number and E^{a} is the blue/redshifted energy of the virtual photon being absorbed by an accelerating charge. E^{a} is determined in the accelerating reference frame N^{a} in which the charge is at rest.

In N^{a}, the frequency of a virtual photon coming from a given direction toward the charge (as seen by the charge) can be written in the vector form

$$f^{\mathrm{a}} = f\left(1 - \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\right) \ , \tag{10.19}$$

where f is the frequency measured at $\boldsymbol{r} = 0$ and $\boldsymbol{r} = \boldsymbol{n}r$, with $\boldsymbol{n}$ a unit vector pointing toward the charge and determining the direction of the incoming virtual photon.

By the uncertainty principle, the energy of a virtual photon of lifetime δt in N^{a} is $\Delta E^{\mathrm{a}} \propto h/\delta t$. A virtual photon of lifetime $\alpha\delta t$ will have energy $\Delta E^{\mathrm{a}}_{\alpha} \propto h/\alpha\delta t = \Delta E^{\mathrm{a}}/\alpha$. By (10.19), the energy of this virtual photon can be written as

$$\Delta E^{\mathrm{a}}_{\alpha} = \frac{\Delta E^{\mathrm{a}}}{\alpha} = \frac{hf^{\mathrm{a}}}{\alpha} = \frac{hf}{\alpha}\left(1 - \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\alpha\right) = \frac{\Delta E}{\alpha}\left(1 - \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\alpha\right) ,$$

where $\Delta E/\alpha = hf/\alpha$ is the energy of that virtual photon determined at $\boldsymbol{r} = 0$. The momentum of the virtual photon will then be

$$\Delta p^{\mathrm{a}}_{\alpha\beta} = \frac{\Delta E^{\mathrm{a}}_{\alpha}}{c}\beta = \frac{\Delta E^{\mathrm{a}}}{c\alpha}\beta = \frac{\Delta E\beta}{c\alpha}\left(1 - \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\alpha\right) .$$

Assume that the number of virtual photons coming from a direction $\boldsymbol{n}$ within the solid angle $\mathrm{d}\Omega$ which are absorbed during the characteristic time δt is x. The momentum of all virtual photons x is then

$$\sum_{i=1}^{x} \Delta p^{\mathrm{a}}_{\alpha_i\beta_i}\boldsymbol{n}\mathrm{d}\Omega = \sum_{i=1}^{x} \frac{\Delta E^{\mathrm{a}}\beta_i}{c\alpha_i}\boldsymbol{n}\mathrm{d}\Omega = \sum_{i=1}^{x} \frac{\Delta E\beta_i}{c\alpha_i}\left(1 - \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\alpha_i\right)\boldsymbol{n}\mathrm{d}\Omega .$$

The force produced by the recoils from all x virtual photons is

$$\mathrm{d}\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = \sum_{i=1}^{x} \frac{\Delta p^{\mathrm{a}}_{\alpha_i\beta_i}}{\delta\tau}\boldsymbol{n}\mathrm{d}\Omega .$$

Then the self-force that acts on the charge and results from the unbalanced recoils of all virtual photons absorbed during the characteristic time δt and coming from all directions toward the charge is

$$\begin{aligned}\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} &= \int \sum_{i=1}^{x} \frac{\Delta p^{\mathrm{a}}_{\alpha_i\beta_i}}{\delta\tau}\boldsymbol{n}\mathrm{d}\Omega = \int \sum_{i=1}^{x} \frac{\Delta E\beta_i}{c\delta\tau\alpha_i}\left(1 - \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{c^2}\alpha_i\right)\boldsymbol{n}\mathrm{d}\Omega \\ &= \int \sum_{i=1}^{x} \frac{\Delta E\beta_i}{c\delta\tau\alpha_i}\boldsymbol{n}\mathrm{d}\Omega - \int \sum_{i=1}^{x} \frac{\Delta E\beta_i}{c^3\delta\tau}\left(\boldsymbol{a}\cdot\boldsymbol{r}\right)\boldsymbol{n}\mathrm{d}\Omega . \qquad (10.20)\end{aligned}$$

Due to symmetry, the first integral in (10.20) is zero. Noting that $\boldsymbol{r} = \boldsymbol{n}r$ and $r = c\delta t$, for the self-force, we can write

$$\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = -\sum_{i=1}^{x} \frac{\Delta E\beta_i\delta t}{c^2\delta\tau}\int \left(\boldsymbol{a}\cdot\boldsymbol{n}\right)\boldsymbol{n}\mathrm{d}\Omega . \qquad (10.21)$$

The integral in (10.21) is similar to the one calculated in Appendix B:

$$\int \left(\boldsymbol{a}\cdot\boldsymbol{n}\right)\boldsymbol{n}\mathrm{d}\Omega = \frac{4\pi}{3}\boldsymbol{a} .$$

Substituting this result into (10.21) and taking into account the fact that

$$U = 4\pi \sum_{i=1}^{x} \Delta E \beta_i$$

is the total energy of all virtual photons (coming from all directions of solid angle 4π) absorbed during the time δt, we obtain for the self-force

$$\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = -\frac{1}{3}\frac{\delta t}{\delta \tau}\frac{U}{c^2}\boldsymbol{a} \ .$$

The energy U represents the energy of the field of the charge. Therefore the quantity U/c^2 is the mass corresponding to the energy of the electric field of the charge and can be regarded as its electromagnetic mass $m^{\mathrm{a}} = U/c^2$.

Let us now consider the contributions from the electromagnetic, weak, and strong interactions to the self-force acting on a particle that participates in the three interactions. The self-force in this case will be

$$\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = -m^{\mathrm{a}}\boldsymbol{a} \ , \tag{10.22}$$

where

$$m^{\mathrm{a}} = \frac{1}{3}\frac{\delta t^{\mathrm{E}}}{\delta \tau^{\mathrm{E}}}\frac{U^{\mathrm{E}}}{c^2} + \frac{1}{3}\frac{\delta t^{\mathrm{W}}}{\delta \tau^{\mathrm{W}}}\frac{U^{\mathrm{W}}}{c^2} + \frac{1}{3}\frac{\delta t^{\mathrm{S}}}{\delta \tau^{\mathrm{S}}}\frac{U^{\mathrm{S}}}{c^2}$$

can be regarded as the inertial mass of the particle which contains contributions from the electromagnetic, weak,[8] and strong interactions. In the expression for m^{a}, the second and third terms contain the energies U^{W} and U^{S} of the weak and strong fields and the times δt^{W}, $\delta \tau^{\mathrm{W}}$, δt^{S}, and $\delta \tau^{\mathrm{S}}$ corresponding to the weak and strong interactions.

Despite the fact that the calculations are semiclassical, everything in the self-force (10.22) shows that it can be regarded as the inertial force to which an accelerating charge is subjected – it is proportional to the acceleration with the correct sign and the coefficient of proportionality has the dimensions of mass. The mass m^{a} in (10.22) is inertial since it is a measure of the resistance the charge offers to its acceleration and is electromagnetic, weak, and strong in origin.

Consider now an electric charge at rest in a gravitational field of strength $\boldsymbol{g}$. According to general relativity, the frequency of an incoming virtual photon which is absorbed by the charge will be shifted:

$$f^{\mathrm{g}} = f\left(1 + \frac{\boldsymbol{g}\cdot\boldsymbol{r}}{c^2}\right) \ , \tag{10.23}$$

[8] For simplicity, the velocity of the carriers of the weak interaction was taken to be equal to c.

where $\boldsymbol{r}$ determines the direction of motion of the incoming virtual photon. If it is approaching the charge from ‘above’ (moving toward the mass producing the gravitational field), the virtual photon will be blueshifted; if it is approaching the charge from ‘below’ (receding from the mass), it will be redshifted.

As in the case of an accelerating charge, the unbalanced recoils from the virtual photons that are absorbed during the time δt by a charge at rest in a gravitational field give rise to a self-force

$$\boldsymbol{F}_{\text{self}}^{\text{g}} = \frac{1}{3}\frac{\delta t}{\delta \tau}\frac{U}{c^2}\boldsymbol{g} \ .$$

Here

$$U = 4\pi \sum_{i=1}^{x} \Delta E \beta_i$$

is also the total energy of all virtual photons (coming from all directions of solid angle 4π) and absorbed by the non-inertial charge during the time δt.

When the contributions from the electromagnetic, weak, and strong interactions are taken into account, the total self-force becomes

$$\boldsymbol{F}_{\text{self}}^{\text{g}} = m^{\text{g}}\boldsymbol{g} \ , \tag{10.24}$$

where

$$m^{\text{g}} = \frac{1}{3}\frac{\delta t^{\text{E}}}{\delta \tau^{\text{E}}}\frac{U^{\text{E}}}{c^2} + \frac{1}{3}\frac{\delta t^{\text{W}}}{\delta \tau^{\text{W}}}\frac{U^{\text{W}}}{c^2} + \frac{1}{3}\frac{\delta t^{\text{S}}}{\delta \tau^{\text{S}}}\frac{U^{\text{S}}}{c^2}$$

is interpreted as the passive gravitational mass of the particle involved in electromagnetic, weak, and strong interactions.

Everything in the self-force (10.24) indicates that it can be regarded as the gravitational force acting on a particle supported in a gravitational field and participating in the three interactions – it is proportional to the gravitational acceleration with the correct sign and the coefficient of proportionality has the dimensions of mass. It is evident that the self-force (10.24) is inertial – it arises only when the particle is prevented from following a geodesic path (i.e., when it is prevented from falling in the gravitational field); only in this case will the particle be subjected to the unbalanced recoils from the incoming virtual quanta whose frequencies are shifted as shown in (10.23). The mass m^{g} in (10.24) is traditionally regarded as passive gravitational mass, but since it is the measure of the resistance the charge offers when deviated from its geodesic path, m^{g} is obviously inertial and electromagnetic, weak, and strong in origin. That is why the masses m^{a} and m^{g} coincide – they are simply the *same* thing:

- the measure of the resistance a particle offers when prevented from following a geodesic path,
- the mass that corresponds to the particle's fields.

When a particle falls in a gravitational field, the frequencies of the incoming virtual quanta are not shifted, as seen by the particle. The recoils from the absorbed quanta cancel out and the particle moves in a non-resistant manner – its worldtube is geodesic. So what makes the worldline of a particle geodesic in terms of the Standard Model is the complete cancellation of the recoils from the incoming virtual quanta that the particle suffers.

The equations (10.22) and (10.24) have the form of Newton's second law. On the one hand, it is clear that in the Standard Model the behaviour of a particle is not governed by the deterministic Newton's second law. On the other hand, however, a non-inertial particle in the Standard Model is also subjected to an inertial or gravitational force which should have the form of Newton's second law.

In this section, we have attempted to explain the origin of inertia in terms of unbalanced recoils from absorbed virtual quanta. The origin of inertia, however, cannot be fully explained without answering the question of the origin of those recoils – whether their very existence implies that they are caused by some kind of inertia of the virtual quanta themselves. Although there have been attempts to attribute inertia to the normal photon [112,113], the issue needs a careful study. But even if it turns out that the virtual quanta do possess inertia, the mechanism examined in the framework of the Standard Model will still account for at least the macroscopic manifestations of inertia.

10.5 Summary

The major issue discussed in this chapter is whether inertia is another manifestation of the reality of spacetime. If the worldtube of a particle is a real four-dimensional object, it follows that it should resist its deformation. As the worldtube of an accelerating particle is deformed (non-geodesic), the resistance the particle offers to its acceleration appears to originate from a four-dimensional stress that arises in the deformed worldtube of the particle. This stress, which is caused by the displacements of the constituents of the accelerated particle from their equilibrium positions, gives rise to a restoring force that tries to bring all constituents back to their non-accelerated positions.

The displacement mechanism has been examined in the case of the classical electron and it has been found that the resulting self-force acting on the classical electron when it accelerates or is supported in a gravitational field does have the form of the inertial force. The same displacement mechanism turns out to be present in the Standard Model as well. The accepted mechanism of interactions through exchange of virtual quanta in the Standard Model in conjunction with the general relativistic shift in the frequencies of the virtual quanta absorbed by a non-inertial particle lead to acceleration-dependent self-interaction effects which appear to account for the origin of inertia and mass of particles. All interactions the Standard Model deals with – electromagnetic, weak, and strong – contribute to the inertia and mass of the particles involved in these interactions.

It may be argued that the mechanism of inertia studied in this chapter can be described in the ordinary three-dimensional language as well. That is true. The real question, however, is whether inertia would exist in a three-dimensional world. We have not answered this question here. What we have shown is that, if spacetime is real, inertia must exist. One of the manifestations of the four-dimensionality of the world will be the existence of inertia through a mechanism which will be the mechanism examined here.

A Classical Electromagnetic Mass Theory and the Arguments Against It

According to the classical electromagnetic mass theory, it is the *unbalanced* repulsion of the volume elements of the charge of an accelerating electron caused by its distorted field that gives rise to the electron's inertia and inertial mass. Since the electric field of an inertial electron (represented by a straight worldline in flat spacetime) is the Coulomb field, the repulsion of its charge elements cancels out exactly and there is no net force acting on the electron. However, if the electron is accelerated, its field distorts, the balance in the repulsion of its volume elements gets disturbed, and as a result it experiences a net self-force $\boldsymbol{F}_{\text{self}}$ which resists its acceleration – it is this resistance that the classical electromagnetic mass theory regards as the electron's inertia. The self-force opposes the external force that accelerates the electron (i.e., its direction is opposite to the electron's acceleration $\mathbf{a}$) and turns out to be proportional to $\boldsymbol{a}$: $\boldsymbol{F}^{\text{a}}_{\text{self}} = -m^{\text{a}}\boldsymbol{a}$, where the coefficient of proportionality m^{a} represents the inertial mass of the electron and is equal to U/c^2, where U is the energy of the electron field; therefore the electron inertial mass is electromagnetic in origin.

The electromagnetic mass of the classical electron can be calculated by three independent methods [114]:

- Energy-derived electromagnetic mass $m_U = U/c^2$, where U is the field energy of an electron at rest. [When the electron is moving with relativistic velocities v, then $m_U = U/\gamma c^2$, where $\gamma = \left(1 - v^2/c^2\right)^{-1/2}$.]
- Momentum-derived electromagnetic mass $m_p = p/v$, where p is the field momentum when the electron is moving at speed v. (For relativistic velocities $m_p = p/\gamma v$.)
- Self-force-derived electromagnetic mass $m_{\text{s}} = F_{\text{self}}/a$, where F_{self} is the self-force acting on the electron when it has acceleration a. (For relativistic velocities $m_{\text{s}} = F_{\text{self}}/\gamma^3 a$.)

There have been two arguments against regarding the entire mass of charged particles as electromagnetic in classical (non-quantum) physics:

- There is a factor of 4/3 which appears in the momentum-derived and self-force-derived electromagnetic mass – $m_p = 4m_U/3$ and $m_\mathrm{s} = 4m_U/3$. (The energy-derived electromagnetic mass m_U does not contain that factor.) Obviously, the three types of electromagnetic mass should be equal.
- The inertia and mass of the classical electron originate from the unbalanced mutual repulsion of its 'parts' caused by the distorted electric field of the electron. However, it is not clear what maintains the electron stable since the classical model of the electron describes its charge as uniformly distributed on a spherical shell, and this means that its volume elements tend to blow up since they repel one another.

Feynman considered the 4/3 factor in the electromagnetic mass expression a serious problem since it made the electromagnetic mass theory (yielding an incorrect relation between energy and momentum due to the 4/3 factor) inconsistent with the special theory of relativity [74, p. 28-4]: "It is therefore impossible to get all the mass to be electromagnetic in the way we hoped. It is not a legal theory if we have nothing but electrodynamics." It seems he was unaware that the 4/3 factor which appears in the momentum-derived electromagnetic mass had already been accounted for in the works of Mandel [96], Wilson [97], Pryce [98], Kwal [99], and Rohrlich [100]. (Each of them removed that factor independently from one another.) The self-force-derived electromagnetic mass has been the most difficult to deal with, persistently yielding the factor of 4/3 [114]. By a covariant application of the Hamilton principle in 1921, Fermi [71] first indirectly showed that there was no 4/3 factor in the self-force acting on a charge supported in a gravitational field. In Chap. 10, we saw how the factor of 4/3 is accounted for in the case of an electron at rest in an accelerating reference frame N^a and in a frame N^g at rest in a gravitational field of strength $\boldsymbol{g}$, described in N^a and N^g, respectively. After the 4/3 factor has been removed, the electromagnetic mass theory of the classical electron becomes fully consistent with relativity and the classical electron mass turns out to be purely electromagnetic in origin.

Since its origin a century ago, the electromagnetic mass theory has not been able to explain why the electron is stable (what holds its charge together). This failure has been seen as related to the presence of the 4/3 factor and has been used as evidence against regarding its entire mass as electromagnetic. To account for the 4/3 factor, it had been assumed that part of the electron mass (regarded as mechanical) originated from non-electric forces (known as the Poincaré stresses [91, 92]) which hold the electron charge together. It was the inclusion of those forces in the classical electron model and the resulting mechanical mass that compensated the 4/3 factor by *reducing* the momentum-derived electromagnetic mass from $(4/3)m_U$ to m_U. This demonstrates that the *attractive* non-electric Poincaré forces make a *negative* contribution to the entire electron mass. It turned out, however, that the 4/3 factor was a result of incorrect calculations of the momentum-derived electromagnetic mass as shown by Mandel [96], Wilson [97], Pryce [98], Kwal [99], and Rohrlich [100]. As there remained nothing to be compensated (in terms of mass), if there were some unknown attractive forces responsible for holding the electron charge together, their *negative* contribution (as attractive forces) to the electron mass would reduce it from m_U to $(2/3)m_U$.

Obviously, there are two options in such a situation – either to seek what this time compensates the negative contribution of the Poincaré stresses to the mass or to assume that the hypothesis of their existence was not necessary in the first place (especially after it turned out that the 4/3 factor does not appear in the correct calculation of the momentum-derived electromagnetic mass). A strong argument supporting the latter option is the fact that, if there existed a real problem with the stability of the electron, the hypothesis of the Poincaré stresses would be needed to balance the mutual repulsion of the volume elements not only of an electron moving with constant velocity (as in the case of the momentum-derived electromagnetic mass), but also of an electron at rest in its rest frame. This, however, is not the case since when the electron is at rest, there is no 4/3 factor problem in the *rest*-energy-derived electromagnetic mass of the electron. If the electron charge tended to blow up as a result of the mutual repulsion of its 'parts', it should do so not only when it is moving at constant velocity but also when it as at rest. Somehow this obvious argument has been overlooked.

Another indication that the stability problem does not appear to be a real problem is that it does not show up (through the 4/3 factor) in the correct calculations of the self-force either. As Fermi [71] showed, and as we have seen in Chap. 10, the Poincaré stresses are not needed for the derivation of the self-force-derived electromagnetic mass since the 4/3 factor which was present in previous derivations of the self-force turned out to be a result of not including in the calculations an anisotropic volume element which arises due to the anisotropic velocity of light in the non-inertial reference frames where the self-force is calculated.

All this implies that there is no real problem with the stability of the electron. We do not know why. What we do know, however, is that, if there were a stability problem, it would inevitably show up in *all* calculations of the energy-derived, momentum-derived, and self-force-derived electromagnetic mass, which is not the case. Obviously, there should be an answer to the question of why calculations based on the indisputably wrong classical model of the electron (i) correctly describe its inertial and gravitational behaviour (including the equivalence of its inertial and passive gravitational mass), and (ii) yield the correct expressions for the inertial and gravitational force. My guess of what that answer might be is that there is something in the classical model which leads to the correct results. Most probably, the spherical distribution of the charge. However, it is not necessary to assume that this distribution is a *solid* spherical shell existing at every single instant as a solid shell (like the *macroscopic* objects we are aware of) – only in that case its 'parts' will repel one another and the sphere will tend to explode. As we have seen in Chap. 6, it is not unthinkable to picture an *elementary* charge as being spherical but not solid (with a continuous distribution of the charge).

The fact that the 4/3 factor has been accounted for and the stability problem does not appear to be a real problem (since it does not show up either in the *rest*-energy-derived electromagnetic mass of the electron, or in the calculations of the self-force) indicates that, in the case of the classical electron, the arguments against regarding its inertia and inertial mass as entirely electromagnetic in origin are answered.

B Calculation of the Self-Force

The self-force

$$\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = \frac{1}{4\pi\epsilon_0}\int\!\!\int\left(\frac{\boldsymbol{n}}{r^2} + \frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{1}{c^2 r}\boldsymbol{a}\right)\left(1 + \frac{\boldsymbol{a}\cdot\boldsymbol{r}}{2c^2}\right)\rho^2 \mathrm{d}V\mathrm{d}V_1$$

can be written (to within terms proportional to c^{-2}) as

$$\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = \frac{1}{4\pi\epsilon_0}\int\!\!\int\left(\frac{\boldsymbol{n}}{r^2} + \frac{3}{2}\frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{1}{c^2 r}\boldsymbol{a}\right)\rho^2 \mathrm{d}V\mathrm{d}V_1 \ . \tag{B.1}$$

We have reached this result assuming that the charge element $\mathrm{d}e^{\mathrm{a}}$ acts upon the charge element $\mathrm{d}e^{\mathrm{a}}_1$. In this case the vector $\boldsymbol{r}$ begins at $\mathrm{d}e^{\mathrm{a}}$ and ends at $\mathrm{d}e^{\mathrm{a}}_1$, i.e., $\boldsymbol{n}$ points from $\mathrm{d}e^{\mathrm{a}}$ to $\mathrm{d}e^{\mathrm{a}}_1$. If we assumed that $\mathrm{d}e^{\mathrm{a}}_1$ acted upon $\mathrm{d}e^{\mathrm{a}}$, the result should be the same. As interchanging the two charge elements reverses the direction of $\boldsymbol{n}$, the self-force in this case will be

$$\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = \frac{1}{4\pi\epsilon_0}\int\!\!\int\left(-\frac{\boldsymbol{n}}{r^2} + \frac{3}{2}\frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{1}{c^2 r}\boldsymbol{a}\right)\rho^2 \mathrm{d}V\mathrm{d}V_1 \ . \tag{B.2}$$

Adding equations (B.2) and (B.1) and dividing the result by 2, we get

$$\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = \frac{1}{4\pi\epsilon_0}\int\!\!\int\left(\frac{3}{2}\frac{\boldsymbol{a}\cdot\boldsymbol{n}}{c^2 r}\boldsymbol{n} - \frac{1}{c^2 r}\boldsymbol{a}\right)\rho^2 \mathrm{d}V\mathrm{d}V_1 \ . \tag{B.3}$$

In order to do the integral (B.3), let us consider the integral [110]

$$\boldsymbol{I} = \int\!\!\int\left(\frac{\boldsymbol{a}\cdot\boldsymbol{n}}{r}\boldsymbol{n}\right)\mathrm{d}V\mathrm{d}V_1 \ . \tag{B.4}$$

We can put $\boldsymbol{n} = \boldsymbol{n}_{\|} + \boldsymbol{n}_{\perp}$, where $\boldsymbol{n}_{\|}$ is parallel to $\boldsymbol{a}$ and $\boldsymbol{n}_{\perp}$ is perpendicular to $\boldsymbol{a}$. Then

$$
\begin{aligned}
(\boldsymbol{a}\cdot\boldsymbol{n})\boldsymbol{n} &= \boldsymbol{a}\cdot\left(\boldsymbol{n}_{\|}+\boldsymbol{n}_{\perp}\right)\left(\boldsymbol{n}_{\|}+\boldsymbol{n}_{\perp}\right)\\
&= \left(\boldsymbol{a}\cdot\boldsymbol{n}_{\|}+\boldsymbol{a}\cdot\boldsymbol{n}_{\perp}\right)\left(\boldsymbol{n}_{\|}+\boldsymbol{n}_{\perp}\right)\\
&= \left(\boldsymbol{a}\cdot\boldsymbol{n}_{\|}\right)\boldsymbol{n}_{\|}+\left(\boldsymbol{a}\cdot\boldsymbol{n}_{\|}\right)\boldsymbol{n}_{\perp}+\left(\boldsymbol{a}\cdot\boldsymbol{n}_{\perp}\right)\boldsymbol{n}_{\|}+\left(\boldsymbol{a}\cdot\boldsymbol{n}_{\perp}\right)\boldsymbol{n}_{\perp}\\
&= \left(\boldsymbol{a}\cdot\boldsymbol{n}_{\|}\right)\boldsymbol{n}_{\|}+\left(\boldsymbol{a}\cdot\boldsymbol{n}_{\|}\right)\boldsymbol{n}_{\perp}\ ,
\end{aligned}
$$

since $(\boldsymbol{a}\cdot\boldsymbol{n}_{\perp}) = 0$. Substituting this result in (B.4) yields

$$
\boldsymbol{I} = \int\int\left(\frac{\boldsymbol{a}\cdot\boldsymbol{n}_{\|}}{r}\boldsymbol{n}_{\|}\right)\mathrm{d}V\mathrm{d}V_1 + \int\int\left(\frac{\boldsymbol{a}\cdot\boldsymbol{n}_{\|}}{r}\boldsymbol{n}_{\perp}\right)\mathrm{d}V\mathrm{d}V_1\ . \qquad \text{(B.5)}
$$

To facilitate the calculations further, let us assume that $\boldsymbol{r}$ is rotated 180° about an axis parallel to $\boldsymbol{a}$ running through the centre of the spherical charge distribution of the electron. Then the vector $\boldsymbol{n} = \boldsymbol{n}_{\|}+\boldsymbol{n}_{\perp}$ becomes $\boldsymbol{n}_{\|}-\boldsymbol{n}_{\perp}$. This means that, in the second integral in (B.5), for every elementary contribution

$$
\left(\frac{\boldsymbol{a}\cdot\boldsymbol{n}_{\|}}{r}\boldsymbol{n}_{\perp}\right)\mathrm{d}V\mathrm{d}V_1\ ,
$$

there is also an equal and opposite contribution

$$
-\left(\frac{\boldsymbol{a}\cdot\boldsymbol{n}_{\|}}{r}\boldsymbol{n}_{\perp}\right)\mathrm{d}V\mathrm{d}V_1\ ,
$$

which shows that the second integral in (B.5) is zero and we can write

$$
\boldsymbol{I} = \int\int\left(\frac{\boldsymbol{a}\cdot\boldsymbol{n}_{\|}}{r}\boldsymbol{n}_{\|}\right)\mathrm{d}V\mathrm{d}V_1\ . \qquad \text{(B.6)}
$$

The integral $\boldsymbol{I}$ is now a function of $\boldsymbol{n}_{\|}$ alone. In order to return to the general case of $\boldsymbol{n}$ (and not restrict ourselves to using $\boldsymbol{n}_{\|}$), we will express the integral in (B.6) in terms of $\boldsymbol{n}$ and a unit vector $\boldsymbol{u}$ in the direction of $\boldsymbol{a}$. Since $\boldsymbol{n}_{\|}$ is parallel to $\boldsymbol{a}$, we have $\boldsymbol{a}\cdot\boldsymbol{n}_{\|} = an_{\|}$. Then we can write

$$
\begin{aligned}
(an_{\|})\boldsymbol{n}_{\|} = \boldsymbol{a}(n_{\|})^2 &= \boldsymbol{a}\left[1^2\left(n_{\|}\right)^2\right]\\
&= \boldsymbol{a}\left(1n_{\|}\right)^2 = \boldsymbol{a}\left(un_{\|}\right)^2 = \boldsymbol{a}(un\cos\theta)^2\\
&= \boldsymbol{a}(\boldsymbol{u}\cdot\boldsymbol{n})^2\ ,
\end{aligned}
$$

where θ is the angle the vector $\boldsymbol{n}$ forms with the acceleration vector $\boldsymbol{a}$. Now we can write the integral (B.6) in the form

$$\boldsymbol{I} = \boldsymbol{a}\int\int \frac{(\boldsymbol{u}\cdot\boldsymbol{n})^2}{r}\mathrm{d}V\mathrm{d}V_1\ . \tag{B.7}$$

Following Abraham [88] and Lorentz [90], we have assumed a spherically symmetric distribution of the electron charge. This shows that, as all directions in space are indistinguishable, the integral in (B.7) should be independent of the direction of the unit vector $\boldsymbol{u}$. Hence the average of this integral over all possible directions of $\boldsymbol{u}$ should be equal to the integral itself:

$$\begin{aligned}\int\int \frac{(\boldsymbol{u}\cdot\boldsymbol{n})^2}{r}\mathrm{d}V\mathrm{d}V_1 &= \frac{1}{4\pi}\int \mathrm{d}\varOmega\int\int \frac{(\boldsymbol{u}\cdot\boldsymbol{n})^2}{r}\mathrm{d}V\mathrm{d}V_1 \\ &= \frac{1}{4\pi}\int\int\frac{\mathrm{d}V\mathrm{d}V_1}{r}\int(\boldsymbol{u}\cdot\boldsymbol{n})^2\mathrm{d}\varOmega\ ,\end{aligned} \tag{B.8}$$

where $\mathrm{d}\varOmega$ is the element of solid angle within which a given unit vector $\boldsymbol{u}$ lies. To do this integral, we choose a polar coordinate system with polar axis along $\boldsymbol{n}$. Then $\boldsymbol{u}\cdot\boldsymbol{n} = \cos\theta$ and $\mathrm{d}\varOmega = \sin\theta\mathrm{d}\theta\mathrm{d}\varphi$ and

$$\begin{aligned}\frac{1}{4\pi}\int(\boldsymbol{u}\cdot\boldsymbol{n})^2\mathrm{d}\varOmega &= \frac{1}{4\pi}\int_0^{\pi}\cos^2\theta\sin\theta\mathrm{d}\theta\int_0^{2\pi}\mathrm{d}\varphi \\ &= \frac{1}{2}\int_0^{\pi}\cos^2\theta\sin\theta\mathrm{d}\theta \\ &= \frac{1}{2}\left(-\frac{1}{3}\cos^3\theta\Big|_0^{\pi}\right) \\ &= \frac{1}{2}\left[-\frac{1}{3}(-1-1)\right] \\ &= \frac{1}{3}\ .\end{aligned}$$

Substituting this result into (B.8) yields

$$\int\int \frac{(\boldsymbol{u}\cdot\boldsymbol{n})^2}{r}\mathrm{d}V\mathrm{d}V_1 = \frac{1}{3}\int\int\frac{\mathrm{d}V\mathrm{d}V_1}{r}\ .$$

Thus for the integral (B.7), we have

$$\boldsymbol{I} = \boldsymbol{a}\int\int \frac{(\boldsymbol{u}\cdot\boldsymbol{n})^2}{r}\mathrm{d}V\mathrm{d}V_1 = \frac{\boldsymbol{a}}{3}\int\int\frac{\mathrm{d}V\mathrm{d}V_1}{r}\ . \tag{B.9}$$

Substituting (B.9) into (B.3), we obtain

$$\begin{aligned}\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} &= \frac{1}{4\pi\epsilon_0}\int\int\left(\frac{3}{2}\frac{\boldsymbol{a}}{3c^2r} - \frac{\boldsymbol{a}}{c^2r}\right)\rho^2\mathrm{d}V\mathrm{d}V_1 \\ &= -\frac{\boldsymbol{a}}{8\pi\epsilon_0c^2}\int\int\frac{\rho^2}{r}\mathrm{d}V\mathrm{d}V_1 \ ,\end{aligned}$$

and finally the expression for the self-force becomes

$$\boldsymbol{F}^{\mathrm{a}}_{\mathrm{self}} = -\frac{U}{c^2}\boldsymbol{a} \ .$$

References

1. H. Weyl: *Philosophy of Mathematics and Natural Science* (Princeton University Press, Princeton 1949) p. 116
2. N. Copernicus: *On the Revolutions of the Heavenly Spheres*. In: *Great Books of the Western World*, Vol. 15, ed. by M.J. Adler (Encyclopedia Britannica, Chicago 1993)
3. J. Kepler: *Harmonies of the World*, Book Five. In: *On the Shoulders of Giants: The Great Works of Physics and Astronomy*, ed. by S. Hawking (Running Press, Philadelphia, London 2002)
4. G. Galileo: *Dialogue Concerning the Two Chief World Systems – Ptolemaic and Copernican*, 2nd edn. (University of California Press, Berkeley 1967)
5. J. Barnes: *Early Greek Philosophy*, 2nd edn. (Penguin Books, London 2001) Part II
6. J. Barnes: *The Presocratic Philosophers* (Routledge, London, New York 1982) Chap. X
7. Aristotle: *Physics*. In: *Great Books of the Western World*, Vol. 7, ed. by M.J. Adler (Encyclopedia Britannica, Chicago 1993)
8. C. Ptolemy: *The Almagest*. In: *Great Books of the Western World*, Vol. 15, ed. by M.J. Adler (Encyclopedia Britannica, Chicago 1993) pp. 8–13
9. H. Minkowski: Space and Time. In: [10] pp. 75–91
10. H.A. Lorentz, A. Einstein, H. Minkowski, and H. Weyl: *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity* (Dover, New York 1952)
11. A. Einstein: On the Electrodynamics of Moving Bodies. In: [10] pp. 37–65
12. A. Sommerfeld: Notes on Minkowski's paper: Space and Time. In: [10] pp. 92–96
13. W. Pauli: *Theory of Relativity* (Dover, New York 1958) p. 4
14. A.P. French: *Special Relativity* (Norton, New York, London 1968) p. 72
15. R. Resnick, D. Halliday: *Basic Concepts in Relativity and Early Quantum Theory*, 2nd edn. (Macmillan, New York 1992) p. 27
16. U.E. Schröder: *Special Relativity* (World Scientific, Singapore 1990) p. 19

17. W.D. McComb: *Dynamics and Relativity* (Oxford University Press, Oxford, New York 1999) p. 187
18. C.H. Hinton: *What is the Fourth Dimension?* (W.S. Sonnenschein and Co., London 1884)
19. C.H. Hinton: *Speculations on the Fourth Dimension: Selected Writings* (Dover, New York 1980)
20. R. Le Poidevin: *Travels in Four Dimensions: The Enigmas of Space and Time* (Oxford University Press, Oxford, New York 2003) p. 48
21. E.F. Taylor, J.A. Wheeler: *Spacetime Physics: Introduction to Special Relativity*, 2nd edn. (Freeman, New York 1992)
22. W. Rindler: *Relativity* (Oxford University Press, Oxford, New York 2001)
23. H. Stein: Phil. Sci. **58**, 147 (1991)
24. C.W. Rietdijk: Phil. Sci. **33**, 341 (1966)
25. H. Putnam: J. Phil. **64**, 240 (1967)
26. N. Maxwell: Phil. Sci. **52**, 23 (1985)
27. H. Stein: J. Phil. **65**, 5 (1968)
28. D. Dieks: Phil. Sci. **55**, 456 (1988)
29. S. McCall and E.J. Lowe: Analysis **63**, 114 (2003)
30. Y. Balashov: On Stages, Worms, and Relativity. In: [32]
31. C. Callender: Phil. Sci. **67** (Proceedings), S587 (2000)
32. C. Callender (Ed.): *Time, Reality and Experience* (Cambridge University Press, Cambridge 2002)
33. S. Saunders: How Relativity Contradicts Presentism. In: [32]
34. S. Savitt: Phil. Sci. **67** (Proceedings), S563 (2000)
35. T. Sider: *Four-Dimensionalism. An Ontology of Persistence and Time* (Clarendon Press, Oxford 2001)
36. B. Rossi, D.B. Hall: Phys. Rev. **57**, 223 (1941)
37. J.B. Hartle: *Gravity: An Introduction to Einstein's General Relativity* (Addison Wesley, San Francisco 2003)
38. J. S. Bell: *Speakable and Unspeakable in Quantum Mechanics* (Cambridge University Press, Cambridge 1987) p. 67
39. K. Gödel: A Remark about the Relationship Between Relativity and Idealistic Philosophy. In: *Albert Einstein: Philosopher–Scientist*, ed. by P. Schilpp (Open Court, La Salle 1949) p. 558
40. V. Petkov: The Flow of Time According to Eleatic Philosophy and the Theory of Relativity. In: *Structur und Dynamik wissenschaftlicher Theorien*, ed. by C. Toegel (P. Lang, Frankfurm am Main Bern New York 1986) pp. 121–149
41. C.W. Misner, K.S. Thorne, J.A. Wheeler: *Gravitation* (Freeman, San Francisco 1973)
42. R.A. Mould, *Basic Relativity* (Springer, Berlin, Heidelberg, New York 1994)

43. R. d'Inverno: *Introducing Einstein's Relativity* (Clarendon Press, Oxford 1992)
44. G.L. Naber: *The Geometry of Minkowski Spacetime* (Springer, Berlin, Heidelberg, New York 1992)
45. P. Kroes: Phil. Sci. **50**, 159–163 (1983)
46. R. Weingard: Brit. J. Phil. Sci. **23**, 119 (1972)
47. V. Petkov: Brit. J. Phil. Sci. **40**, 69 (1989)
48. D. Dieks: Space, Time and Coordinates in a Rotating World. In: *Relativity in Rotating Frames: Relativistic Physics in Rotating Reference Frames* (Fundamental Theories of Physics, Vol. 135), ed. by G. Rizzi, M.L. Ruggiero (Kluwer, Dordrecht, Boston, London 2004) pp. 29–42
49. Aristotle: *Physics*, Book IV, Chap. 14. In: *Great Books of the Western World*, Vol. 7, ed. by M.J. Adler (Encyclopedia Britannica, Chicago 1993)
50. Saint Augustine: *The Confessions*, Book XI. In: *Great Books of the Western World*, Vol. 16, ed. by M.J. Adler (Encyclopedia Britannica, Chicago 1993)
51. V. Petkov: Weyl's View on the Objective World. In: *Exact Sciences and Their Philosophical Foundations*, ed. by W. Deppert, K. Huebner, A. Oberschelp, V. Weidemann (P. Lang, Frankfurm am Main Bern New York, Paris 1988) pp. 519–524
52. A.H. Anastassov: The Theory of Relativity and the Quantum Action (4-Atomism),DSc Thesis, Sofia University, Sofia (1984); *The Theory of Relativity and the Quantum Action* (Nautilus, Sofia 2003), in Bulgarian
53. A.H. Anastassov: Annuaire de l'Université de Sofia, St. Kliment Ohridski, Faculté de Physique, **81**, 135 (1993)
54. *Encyclopedia of Physics*, 2nd edn, ed. by R.G. Lerner and G.L. Trigg (VCH Publishers, New York 1991) p. 721
55. J. Trefil: *The Nature of Science* (Houghton Mifflin, Boston New York 2003) p. 98
56. B.G. Kuznetsov: *Einstein and Dostoyevsky* (Hutchinson, London 1972)
57. H. Ohanian, R. Ruffini: *Gravitation and Spacetime*, 2nd edn. (W.W. Norton, New York, London 1994) p. 197
58. A. Einstein, L. Infeld: *The Evolution of Physics* (Simon and Schuster, New York 1966) p. 221
59. P.A. Tipler: *Physics*, Vol. 3, 4th edn. (Freeman, New York 1999) p. 1272
60. R.L. Reese: *University Physics*, Vol. 2 (Brooks/Cole, New York 2000) p. 1191
61. R.A. Serway: *Physics*, Vol. 2, 4th edn. (Saunders, Chicago 1996) p. 1180
62. P.M. Fishbane, S. Gasiorowicz, S.T. Thornton: *Physics* (Prentice Hall, New Jersey 1993) p. 1192
63. G. Rizzi, M.L. Ruggiero (Eds.): *Relativity in a Rotating Frame* (Kluwer, Dordrecht 2003)

64. I.I. Shapiro: Phys. Rev. Lett. **13**, 789 (1964)
65. I.I. Shapiro: Phys. Rev. **141**, 1219 (1966)
66. E.F. Taylor, J.A. Wheeler: *Exploring Black Holes: Introduction to General Relativity* (Addison Wesley Longman, San Francisco 2000) p. E-1
67. A. Harpaz, N. Soker: Gen. Rel. Grav. **30**, 1217 (1998); see also physics/9910019
68. W. Rindler: Am. J. Phys. **36**, 540 (1968)
69. A. Einstein: Ann. Phys. **49** (1916). In: [10] pp. 111–164
70. L. Lerner: Am. J. Phys. **65**, 1194 (1997)
71. E. Fermi: Nuovo Cimento **22**, 176 (1921)
72. D.J. Griffiths: *Introduction to Electrodynamics*, 2nd edn. (Prentice Hall, London 1989) p. 416
73. W.K.H. Panofsky and M. Phillips: *Classical Electricity and Magnetism*, 2nd edn. (Addison-Wesley, Massachusetts, London 1962) p. 342
74. R.P. Feynman, R.B. Leighton and M. Sands: *The Feynman Lectures on Physics*, Vol. 2 (Addison-Wesley, New York 1964) p. 21-10
75. M. Schwartz: *Principles of Electrodynamics* (Dover, New York 1972) p. 213
76. J.D. Jackson, *Classical Electrodynamics*, 3rd edn. (Wiley, New York 1999) p. 664
77. F. Rohrlich: *Classical Charged Particles* (Addison-Wesley, New York 1990) p. 218
78. M. von Laue: *Relativitäts Theorie*, 3rd edn., Vol. 1 (Frederick Vieweg und Sohn, Braunschweig 1919)
79. T. Fulton, F. Rohrlich: Ann. Phys. **9,** 499 (1960)
80. F. Rohrlich: Ann. Phys. **22**, 169 (1963)
81. D.G. Boulware: Ann. Phys. **124**, 169 (1980)
82. B.S. DeWitt and R.W. Brehme: Ann. Phys. **9**, 220 (1960)
83. W. Rindler: *Essential Relativity* (Springer, Berlin, Heidelberg, New York 1997) p. 244
84. J.L. Synge: *Relativity: The General Theory* (Nord-Holland, Amsterdam 1960) p. 109
85. J.J. Thomson: Phil. Mag. **11**, 229 (1881)
86. O. Heaviside: *The Electrician* **14**, 220 (1885)
87. G.F.C. Searle: Phil. Mag. **44**, 329 (1897)
88. M. Abraham: *The Classical Theory of Electricity and Magnetism*, 2nd edn. (Blackie, London 1950)
89. H.A. Lorentz: Proceedings of the Academy of Sciences of Amsterdam **6**, 809 (1904)
90. H.A. Lorentz: *Theory of Electrons*, 2nd edn. (Dover, New York 1952)
91. H. Poincaré: Compt. Rend. **140**, 1504 (1905)
92. H. Poincaré: Rendiconti del Circolo Matematico di Palermo **21**, 129 (1906)
93. E. Fermi: Phys. Zeits. **23**, 340 (1922)

94. E. Fermi: Rend. Acc. Lincei (5) **31**, 184; 306 (1922)
95. E. Fermi: Nuovo Cimento **25**, 159 (1923)
96. H. Mandel: Z. Physik **39**, 40 (1926)
97. W. Wilson: Proc. Phys. Soc. **48**, 736 (1936)
98. M.H.L. Pryce: Proc. Roy. Soc. A **168**, 389 (1938)
99. B. Kwal: J. Phys. Rad. **10**, 103 (1949)
100. F. Rohrlich: Am. J. Phys. **28**, 639 (1960)
101. M. Jammer: *Concepts of Mass in Classical and Modern Physics* (Dover, New York, 1997), Chap. 11. See also [102]
102. M. Jammer: *Concepts of Mass in Contemporary Physics and Philosophy* (Princeton University Press, Princeton, 2000) p. 34
103. J.W. Butler: Am. J. Phys. **37**, 1258 (1969)
104. E. Mach: *Science of Mechanics*, 9th edn. (Open Court, London 1933)
105. P. Pearle: Classical Electron Models. In: *Electromagnetism: Paths to Research*, ed. by D. Teplitz (Plenum Press, New York 1982) pp. 211–295
106. D. Bender et al: Phys. Rev. **D30,** 515 (1984)
107. M.H. MacGregor: *The Enigmatic Electron* (Kluwer, Dordrecht 1992)
108. D. Hestenes, A. Weingartshofer (Eds.): *The Electron: New Theory and Experiment* (Kluwer, Dordrecht 1991)
109. M. Springford (Ed.): *Electron: A Centenary Volume* (Cambridge University Press, Cambridge 1997)
110. B. Podolsky, K.S. Kunz: *Fundamentals of Electrodynamics* (Marcel Dekker, New York 1969) p. 288
111. M.K. Gaillard, P.D. Grannis, and F.J. Sciulli: Rev. Mod. Phys. **71**, No. 2 S96 (1999)
112. A. Einstein: Ann. Phys. **20**, 627 (1906)
113. D.L. Livesey: *Atomic and Nuclear Physics* (Blaisdell, Massachusetts 1966) p. 117
114. D. J. Griffiths and R. E. Owen: Am. J. Phys. **51**, 1120 (1983)

Index