Vesselin Petkov
Editor

FUNDAMENTAL THEORIES OF PHYSICS 165

Minkowski Spacetime: A Hundred Years Later

Springer

Minkowski Spacetime: A Hundred Years Later

Fundamental Theories of Physics

An International Book Series on The Fundamental Theories of Physics: Their Clarification, Development and Application

Volume 165

For other titles published in this series, go to http://www.springer.com/series/6001

Vesselin Petkov
Editor

Minkowski Spacetime: A Hundred Years Later

Springer

Editor
Vesselin Petkov
Science College
Concordia University
1455 de Maisonneuve Blvd. West
Montreal QC H3G 1M8
Canada
vpetkov@alcor.concordia.ca

ISBN 978-90-481-3474-8 e-ISBN 978-90-481-3475-5
DOI 10.1007/978-90-481-3475-5
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009941860

Cover design: eStudio Calamar S.L.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume is dedicated to the one hundredth anniversary of the publication of Hermann Minkowski's paper "Raum und Zeit" in 1909 [1]. The paper presents the text of the talk Minkowski gave at the 80th Meeting of the German Natural Scientists and Physicians in Cologne on September 21, 1908.

Minkowski's work on the spacetime representation of special relativity had a huge impact on the twentieth century physics, which can be best expressed by merely stating what is undeniable – that modern physics would be impossible without the notion of spacetime. It is sufficient to mention as an example only the fact that general relativity would be impossible without this notion; Einstein succeeded to identify gravity with the curvature of spacetime only when he overcame his initial hostile reaction to Minkowski's four-dimensional representation of special relativity and adopted spacetime as the correct relativistic picture of the world.

While there exists an unanimous consensus on the mathematical significance of spacetime for theoretical physics, for a hundred years there has been no consensus on the nature of spacetime itself. The first sign of this continuing controversy was Sommerfeld's remark in his notes on Minkowski's article [2]: "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".

As we owe Minkowski, especially now, a clear answer to the question of the nature of spacetime – whether it is only a mathematical space or represents a real four-dimensional world – I think every physicist, particularly relativists, should read his paper, because even a century after its publication there are physicists who still do not seem to appreciate fully the depth of Minkowski's ideas on space and time and on the *physical meaning* of special relativity. One can often hear that the three-dimensional and the four-dimensional representations of relativity are just different *descriptions* of the relativistic phenomena[1]. Such a position does not reflect the profound meaning of Minkowski's representation of special relativity and as a result does not lead to a genuine understanding of what relativity is telling us

[1] These claims somehow ignore the fact that general relativity cannot be adequately represented in a three-dimensional language.

about the world. The reason is that the two formulations of relativity represent a three-dimensional and a four-dimensional world and, obviously, the answer to the question of what is the dimensionality of the world is not a matter of convenient description.

The issue of whether or not spacetime represents a real four-dimensional world is especially relevant now when we celebrate Minkowski's insight that the profound physical meaning of special relativity is most adequately expressed not by the relativity postulate but by Minkowski's postulate of the absolute world (spacetime). More specifically, the issue of whether or not spacetime is 'just a convenient description' is particularly relevant to the question of why we celebrate Minkowski, not Poincaré, given the fact that it was Poincaré who first realized (before July 1905) that the Lorentz transformations have a natural geometric interpretation as rotations in a four-dimensional space whose fourth dimension is time [10, p.168].

We can now only guess why Poincaré did not develop further this revolutionary idea. The most probable explanation might be his conventionalism[2] – he believed that our physical theories are nothing more than convenient descriptions of the world and therefore it is really a matter of *convenience* which theory we would use in a given situation. So Poincaré appeared to have seen nothing revolutionary in the idea of a mathematical four-dimensional space since such an idea would not necessarily force us to assume that the world itself is also four-dimensional.

By contrast, Minkowski, who almost certainly had been aware of Poincaré's geometric interpretation of the Lorentz transformations, realized that the relativity postulate makes sense only in a *real* four-dimensional world. Had he believed, apparently like Poincaré, that uniting space and time into a four-dimensional space was only a convenient mathematical device, he would not have written a paper whose title and content were devoted to something the main idea of which had already been published by Poincaré two years earlier (and written three years earlier) and would not have begun his paper with the now famous introduction, which unequivocally announced the revolution in our views on space and time [6]:

> The views on space and time which I wish to lay before you have sprung from the soil of experimental physics. Therein lies their strength. Their tendency is 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.

Minkowski clearly realized that "The world-postulate allows identical treatment of the four coordinates x, y, z, t" since these coordinates represent the temporal and

---

[2] The failure of Poincaré to comprehend the profound physical meaning of the principle of relativity and the realized by him geometric interpretation of the Lorentz transformations is one of the examples in the history of physics when an inadequate philosophical position prevents a scientist (even as great as Poincaré) from making a discovery. This is both sad and ironic because physicists often think that they do not need any philosophical position for their research. As Daniel Dennett put it [5]: "Scientists sometimes deceive themselves into thinking that philosophical ideas are only, at best, decorations or parasitic commentaries on the hard, objective triumphs of science, and that they themselves are immune to the confusions that philosophers devote their lives to dissolving. But there is no such thing as philosophy-free science; there is only science whose philosophical baggage is taken on board without examination."

spatial dimensions, which have equal status in spacetime. The identical status of space and time in spacetime, according to the postulate of the absolute world, naturally explains the identical status of all inertial observers as well. The very existence of spacetime[3] makes it possible for the inertial observers to describe physical phenomena in terms of their own times and *spaces*; an inertial observer's time and space are as good as the times and spaces of the other inertial observers exactly as the very existence of a plane makes the x and y axes of a coordinate system as good as the x and y axes of any other coordinate system describing the same plane.

While Einstein, unlike Lorentz, insisted that the times of all inertial observers are equally good, Minkowski noticed that "neither Einstein nor Lorentz made any attack on the concept of space" [6] and stressed that the idea of many spaces is unavoidable in special relativity [6]:

> We would then have in the world no longer *the* 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. You see why I said at the outset that space and time are to fade away into shadows, and that only a world in itself will subsist.

But inertial observers in relative motion can have different spaces, which is implied by the relativity postulate and explicitly following from Minkowski's world-postulate, only in a real four-dimensional world with one temporal and three spatial dimensions. With this in mind, one can now understand why Minkowski saw special relativity as revolutionizing our views on space and time. This can be even better understood by taking into account what Minkowski appeared to have realized – that *special relativity is impossible in a three-dimensional world* because a three-dimensional world entails that there exists *one* absolute space that is common to all inertial observers in relative motion. Therefore, if the world were three-dimensional, all inertial observers would share the *same* absolute space and the *same* class of absolutely simultaneous events (because space is defined in terms of simultaneity) in contradiction with relativity.

That Minkowski took the idea of the four-dimensionality of the world seriously is seen throughout his paper. It is sufficient to mention here only his explanation of Lorentz's contraction hypothesis. According to Minkowski's explanation the length contraction of the Lorentzian electrons is a manifestation of the reality of the electrons' worldtubes (Minkowski used the word 'bands') and therefore of the four-dimensionality of the world: the fact that the spaces of two inertial observers in relative motion intersect the worldtube of a Lorentzian electron in two three-dimensional cross sections of different lengths is possible only in a four-dimensional world, where the two observers can have different spaces and the worldtube of the electron is a real four-dimensional object in order that the observers can regard different cross sections of it as the three-dimensional electron they measure.

[3] By "existence of spacetime" I mean the existence of the four-dimensional world in which time and space are dimensions, not the debate substantivalism (absolutism) versus relationism.

The volume begins with an excellent retranslation of Minkowski's paper by Dennis Lehmkuhl of Oxford University and the original German text of the article. The fourteen contributed papers are divided into three parts entitled "The Impact of Minkowski Spacetime on the Twentieth Century Physics from a Historical Perspective," "Implications of Minkowski Spacetime for Theoretical Physics," and "Conceptual and Philosophical Issues of Minkowski Spacetime."

I would like to thank sincerely Springer and especially Dr. Maria Bellantone and Ms. Mieke van der Fluit for making it possible to include a retranslation of Minkowski's paper in his centennial volume.

Montreal *Vesselin Petkov*
7 May 2009

References

1. H. Minkowski, "Raum und Zeit", *Physikalische Zeitschrift* **10** (1909) pp 104–111; *Jahresbericht der Deutschen Mathematiker-Vereinigung* **18** (1909) pp 75–88
2. A. Sommerfeld, Notes on Minkowski's paper "Space and Time". In: [3] p 92
3. 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)
4. H. Poincaré, "Sur la dynamique de l'électron", *Rendiconti del Circolo matematico Rendiconti del Circolo di Palermo* **21** (1906) pp 129–176
5. D.C. Dennett, *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (Simon and Schuster, New York 1996) p 21
6. H. Minkowski, "Space and Time," this volume

Contents

Contributors

Abhay Ashtekar Department of Physics, Institute for Gravitation and the Cosmos, Pennsylvania State University, Penn State, University Park, PA 16802, USA

Orfeu Bertolami Dpto. Física, Instituto Superior Técnico, Lisbon, Portugal

Martin Bojowald Institute for Gravitational and the Cosmos, The Pennsylvania State University, 104 Davey Lab, University Park, PA 16802, USA

Leo Corry Cohn Institute for History and Philosophy of Science, Tel-Aviv University, Ramat Aviv, 69978, Israel

Dennis Dieks Department of History and Foundations of Science, Utrecht University P.O. Box 80010, NL 3508 TA Utrecht, The Netherlands

Rodolfo Gambini Department of Physics and Astronomy, Louisiana State University. Baton Rouge, LA, 70803-4001, USA

Yvon Gauthier Department of Philosophy, University of Montreal, Montreal, QC H3C 3J7, Canada

Domenico Giulini Max-Planck-Institute for Gravitational Physics, Albert-Einstein-Institute, Institute of Physics, University of Freiburg, Am Mühlenberg, 1, D-14476 Golm/Postdam, Germany

Graham Hall Department of Mathematical Sciences, University of Aberdeen, Scotland, UK

Dennis Lehmkuhl Oriel College, Oxford University and Institute of Philosophy and Interdisciplinary Centre for Science and Technology Studies, University of Wuppertal

Hermann Minkowski Author of the Original German Text "Raum und Zeit"

Vesselin Petkov Science College, Concordia University, Montreal, QC H3G 1M8, Canada

Herbert Pietschmann Institute for Theoretical Physics, University of Vienna, Austria

Jorge Pullin Department of Physics and Astronomy, Louisiana State University Baton Rouge, LA 70803-4001, USA

William G. Unruh Department of Physics and Astronomy, University of British Columbia, Vancouver, Canada V6T 1Z1

Scott Walter Université Nancy 2 and H. Pointcaré Archives (CNSR, UMR 7117), 23 bd Albert Ier, 54015 Nancy, France

Paul S. Wesson Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

H. Minkowski

Raum und Zeit

H. Minkowski[1]

M. H.! Die Anschauungen über Raum und Zeit, die ich Ihnen entwickeln möchte, sind auf experimentell-physikalischem Boden erwachsen. Darin liegt ihre Stärke. Ihre Tendenz ist eine radikale. Von Stund an sollen Raum für sich und Zeit für sich völlig zu Schatten herabsinken und nur noch eine Art Union der beiden soll Selbständigkeit bewahren.

I.

Ich möchte zunächst ausführen, wie man von der gegenwärtig angenommenen Mechanik wohl durch eine rein mathematische Überlegung zu veränderten Ideen über Raum und Zeit kommen könnte. Die Gleichungen der Newtonschen Mechanik zeigen eine zweifache Invarianz. Einmal bleibt ihre Form erhalten, wenn man das zugrunde gelegte räumliche Koordinatensystem einer beliebigen *Lagenveränderung* unterwirft, zweitens, wenn man es in seinem Bewegungszustande verändert, nämlich ihm irgendeine *gleichförmige Translation* aufprägt; auch spielt der Nullpunkt der Zeit keine Rolle. Man ist gewohnt, die Axiome der Geometrie als erledigt anzusehen, wenn man sich reif für die Axiome der Mechanik fühlt, und deshalb werden jene zwei Invarianzen wohl selten in einem Atemzuge genannt. Jede von ihnen bedeutet eine gewisse Gruppe von Transformationen in sich für die Differentialgleichungen der Mechanik. Die Existenz der ersteren Gruppe sieht man als einen fundamentalen Charakter des Raumes an. Die zweite Gruppe straft man am liebsten mit Verachtung, um leichten Sinnes darüber hinwegzukommen, daß man von den physikalischen Erscheinungen her niemals entscheiden kann, ob der als ruhend vorausgesetzte Raum sich nicht am Ende in einer gleichförmigen Translation befindet. So führen jene zwei Gruppen ein völlig getrenntes Dasein nebeneinander. Ihr gänzlich heterogener Charakter mag davon abgeschreckt haben, sie zu komponieren. Aber gerade die komponierte volle Gruppe als Ganzes gibt uns zu denken auf.

[1] Vortrag, gehalten auf der 80. Versammlung Deutscher Naturforscher und Ärzte zu Cöln am 21. September 1908.

Space and Time

H. Minkowski[1]

Gentlemen! The views on space and time which I wish to lay before you have sprung from the soil of experimental physics. Therein lies their strength. Their tendency is 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.⋆

I.

First of all I would like to show how it might be possible, setting out from the adopted mechanics of the present day, along a purely mathematical line of thought, to arrive at changed ideas of space and time. The equations of Newtonian mechanics exhibit a twofold invariance. Their form remains unaltered, firstly, if we subject the assumed system of spatial coordinates to an arbitrary *change of position*; secondly if we change its state of motion, namely by imparting to it any *uniform translation*; furthermore, the zero point of time is given no part to play. We are accustomed to look upon the axioms of geometry as finished with, when we feel ripe for the axioms of mechanics, and it is probably for that reason that these two invariances are rarely mentioned in the same breath. Each of them by itself signifies a certain group of transformations for the differential equations of mechanics. The existence of the first group is looked upon as a fundamental characteristic of space. The second group is preferably treated with disdain, so that one does not have to trouble oneself with the difficulty of never being able to decide, based on physical phenomena, whether space, which is assumed to be stationary, may not be after all in a state of uniform translation. Thus the two groups, side by side, lead their lives entirely apart. Their utterly heterogeneous character may have discouraged any attempt to compose them. But it is precisely the complete, composed group, as a whole, that gives us to think.

[1] Talk, given at the 80th Meeting of German Scientists and Physicians, Cologne, September 21, 1908.

Wir wollen uns die Verhältnisse graphisch zu veranschaulichen suchen. Es seien x, y, z rechtwinklige Koordinaten für den Raum, und t bezeichne die Zeit. Gegenstand unserer Wahrnehmung sind immer nur Orte und Zeiten verbunden. Es hat niemand einen Ort anders bemerkt als zu einer Zeit, eine Zeit anders als an einem Orte. Ich respektiere aber noch das Dogma, daß Raum und Zeit je eine unabhängige Bedeutung haben. Ich will einen Raumpunkt zu einem Zeitpunkt, d. i. ein Wertsystem x, y, z, t einen *Weltpunkt* nennen. Die Mannigfaltigkeit aller denkbaren Wertsysteme x, y, z, t soll die *Welt* heißen. Ich könnte mit kühner Kreide vier Weltachsen auf die Tafel werfen. Schon *eine* gezeichnete Achse besteht aus lauter schwingenden Molekülen und macht zudem die Reise der Erde im All mit, gibt also bereits genug zu abstrahieren auf; die mit der Anzahl 4 verbundene etwas größere Abstraktion tut dem Mathematiker nicht wehe. Um nirgends eine gähnende Leere zu lassen, wollen wir uns vorstellen, daß aller Orten und zu jeder Zeit etwas Wahrnehmbares vorhanden ist. Um nicht Materie oder Elektrizität zu sagen, will ich für dieses Etwas das Wort Substanz brauchen. Wir richten unsere Aufmerksamkeit auf den im Weltpunkt x, y, z, t vorhandenen substantiellen Punkt und stellen uns vor, wir sind imstande, diesen substantiellen Punkt zu jeder anderen Zeit wieder zu erkennen. Einem Zeitelement dt mögen die Änderungen dx, dy, dz der Raumkoordinaten dieses substantiellen Punktes entsprechen. Wir erhalten alsdann als Bild sozusagen für den ewigen Lebenslauf des substantiellen Punktes eine Kurve in der Welt, eine *Weltlinie*, deren Punkte sich eindeutig auf den Parameter t von $-\infty$ bis $+\infty$ beziehen lassen. Die ganze Welt erscheint aufgelöst in solche Weltlinien, und ich möchte sogleich vorwegnehmen, daß meiner Meinung nach die physikalischen Gesetze ihren vollkommensten Ausdruck als Wechselbeziehungen unter diesen Weltlinien finden dürften.

Durch die Begriffe Raum und Zeit fallen die x, y, z-Mannigfaltigkeit $t = 0$ und ihre zwei Seiten $t > 0$ und $t < 0$ auseinander. Halten wir der Einfachheit wegen den Nullpunkt von Raum und Zeit fest, so bedeutet die zuerst genannte Gruppe der Mechanik, daß wir die x, y, z-Achsen in $t = 0$ einer beliebigen Drehung um den Nullpunkt unterwerfen dürfen, entsprechend den homogenen linearen Transformationen des Ausdrucks

$$x^2 + y^2 + z^2$$

in sich. Die zweite Gruppe aber bedeutet, daß wir, ebenfalls ohne den Ausdruck der mechanischen Gesetze zu verändern,

$$x, y, z, t \quad \text{durch} \quad x - \alpha t,\ y - \beta t,\ z - \gamma t, t$$

mit irgendwelchen Konstanten α, β, γ ersetzen dürfen. Der Zeitachse kann hiernach eine völlig beliebige Richtung nach der oberen halben Welt $t > 0$ gegeben werden. Was hat nun die Forderung der Orthogonalitat im Raume mit dieser völligen Freiheit der Zeitachse nach oben hin zu tun?

Die Verbindung herzustellen, nehmen wir einen positiven Parameter c und betrachten das Gebilde

$$c^2 t^2 - x^2 - y^2 - z^2 = 1.$$

We shall aim to visualize the state of the things by the graphic method. Let x, y, z be orthogonal coordinates for space, and let t denote time. The objects of our perception invariably include places and times in combination. Nobody has ever noticed a place except at a time, or a time except at a place. But I respect the dogma that space and time each have an independent meaning. I will call a point of space at a certain point of time, i.e. a system of values x, y, z, t, a *world-point*. The manifold of all conceivable x, y, z, t systems shall be called the *world*. With this most valiant piece of chalk I might cast upon the blackboard four world-axes. Even merely *one* chalky axis consists of a number of oscillating molecules, and moreover is taking part in the earth's travels through the universe, hence already demands of us an ample amount of abstraction; the somewhat greater abstraction connected to the number 4 does not hurt the mathematician. Not to leave a yawning void anywhere, we will imagine that everywhere and everywhen there is something perceptible. To avoid saying matter or electricity, I will use for this something the word substance. We fix our attention on the substantial point which is at the world-point x, y, z, t, and imagine that we are able to recognize this substantial point at any other time. Corresponding to a time element $\mathrm{d}t$, we have the variations $\mathrm{d}x, \mathrm{d}y, \mathrm{d}z$ of the space coordinates of this substantial point. Then we obtain, as an image, so to speak, of the eternal career of the substantial point, a curve in the world, a *world-line*, the points of which can be referred to uniquely to the parameter t running from $-\infty$ to $+\infty$. The whole world seems to resolve itself into such world-lines, and I would fain anticipate myself by saying that in my opinion the laws of physics might find their most perfect expression as interrelations between these world-lines.

Due to the concepts of 'space' and 'time', the x, y, z-manifold $t = 0$ and its two sides $t > 0$ and $t < 0$ fall asunder. If, for simplicity, we retain the same zero point of space and time, the first-mentioned group of mechanics means that we may subject the axes of x, y, z at $t = 0$ to an arbitrary rotation about the origin, corresponding to the homogeneous linear transformations of the expression

$$x^2 + y^2 + z^2.$$

But the second group means that we may – also without changing the expression of the laws of mechanics – replace

$$x, y, z, t \quad \text{by} \quad x - \alpha t, y - \beta t, z - \gamma t, t,$$

where α, β, γ are arbitrary constants. The time axis can hence be given a completely arbitrary direction towards the upper half of the world, $t > 0$. Now what has the requirement of orthogonality in space to do with this complete freedom of the time axis in an upward direction?

To establish the connection, let us take a positive parameter c, and consider the structure

$$c^2t^2 - x^2 - y^2 - z^2 = 1.$$

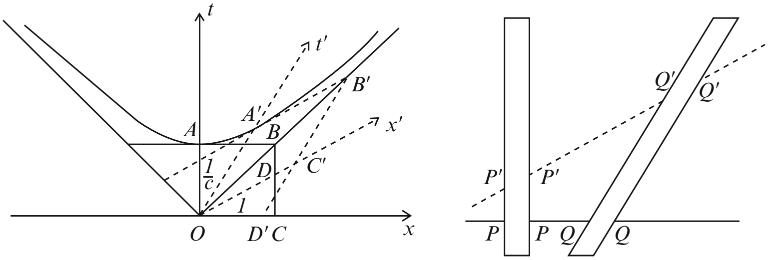


Fig. 1

Es besteht aus zwei durch $t = 0$ getrennten Schalen nach Analogie eines zweischaligen Hyperboloids. Wir betrachten die Schale im Gebiete $t > 0$, und wir fassen jetzt diejenigen homogenen linearen Transformationen von x, y, z, t in vier neue Variable x', y', z', t' auf, wobei der Ausdruck dieser Schale in den neuen Variabeln entsprechend wird. Zu diesen Transformationen gehören offenbar die Drehungen des Raumes um den Nullpunkt. Ein volles Verständnis der übrigen jener Transformationen erhalten wir hernach bereits, wenn wir eine solche unter ihnen ins Auge fassen, bei der y und z ungeändert bleiben. Wir zeichnen (Fig. 1) den Durchschnitt jener Schale mit der Ebene der x- und der t-Achse, den oberen Ast der Hyperbel $c^2t^2 - x^2 = 1$, mit seinen Asymptoten. Ferner werde ein beliebiger Radiusvektor OA' dieses Hyperbelastes vom Nullpunkte O aus eingetragen, die Tangente in A' an die Hyperbel bis zum Schnitte B' mit der Asymptote rechts gelegt, $OA'B'$ zum Parallelogramm $OA'B'C'$ vervollständigt, endlich für das spätere noch $B'C'$ bis zum Schnitt D' mit der x-Achse durchgeführt. Nehmen wir nun OC' und OA' als Achsen für Parallelkoordinaten x', t' mit den Maßstäben $OC' = 1, OA' = 1/c$, so erlangt jener Hyperbelast wieder den Ausdruck $c^2t'^2 - x'^2 = 1, t' > 0$, und der Übergang von x, y, z, t zu x', y, z, t' ist eine der fraglichen Transformationen. Wir nehmen nun zu den charakterisierten Transformationen noch die beliebigen Verschiebungen des Raum- und Zeit-Nullpunktes hinzu und konstituieren damit eine offenbar noch von dem Parameter c abhängige Gruppe von Transformationen, die ich mit G_c bezeichne.

Lassen wir jetzt c ins Unendliche wachsen, also $1/c$ nach Null konvergieren, so leuchtet an der beschriebenen Figur ein, daß der Hyperbelast sich immer mehr der x-Achse anschmiegt, der Asymptotenwinkel sich zu einem gestreckten verbreitert, jene spezielle Transformation in der Grenze sich in eine solche verwandelt, wobei die t'-Achse eine beliebige Richtung nach oben haben kann und x' immer genauer sich an x annähert. Mit Rücksicht hierauf ist klar, daß aus der Gruppe G_c in der Grenze für $c = \infty$, also als Gruppe G_∞, eben jene zu der Newtonschen Mechanik gehörige volle Gruppe wird. Bei dieser Sachlage, und da G_c mathematisch verständlicher ist als G_∞ hätte wohl ein Mathematiker in freier Phantasie auf den Gedanken verfallen können, daß am Ende die Naturerscheinungen tatsächlich eine Invarianz nicht bei der Gruppe G_∞, sondern vielmehr bei einer Gruppe G_c mit bestimmtem endlichen, nur in den gewöhnlichen Maßeinheiten *äußerst großen* c besitzen.

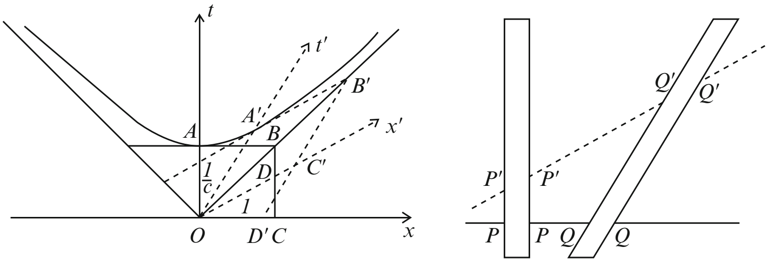


Fig. 1

It consists of two sheets separated by $t = 0$, on the analogy of a hyperboloid of two sheets. We consider the sheet in the region $t > 0$, and now take those linear homogeneous linear transformations of x, y, z, t into four new variables x', y', z', t', where the expression for this sheet in the new variables is of the same form. It is evident that the rotations of space about the origin pertain to these transformations. Thus we gain full comprehension of the rest of the transformations simply by taking into consideration one among them, such that y and z remain unchanged. We draw (Fig. 1) the section of this sheet by the plane of the x- and the t-axis – the upper branch of the hyperbola $c^2t^2 - x^2 = 1$, with its asymptotes. Furthermore, from the origin O we draw an arbitrary radius vector OA' of this branch of the hyperbola; then the tangent to the hyperbola at A' to cut the asymptote on the right at B'; we complete $OA'B'$ to the parallelogram $OA'B'C'$; and finally, for subsequent use, we produce $B'C'$ so that it cuts the axis of x at D'. Now if we take OC' and OA' as axes for alternative coordinates x', t', with the measures $OC' = 1$, $OA' = 1/c$, then that branch of the hyperbola again acquires the expression $ct'^2 - x'^2 = 1, t' > 0$, and the transition from x, y, z, t to x', y, z, t' is one of the transformations in question. To these transformations we now add the arbitrary displacements of the zero point of space and time, and thereby constitute a group of transformations which is evidently still dependent on the parameter c. This group I denote by G_c.

If we now allow c to increase to infinity, and $1/c$ therefore to converge towards zero, it becomes clear from the figure that the branch of the hyperbola bends more and more towards the x axis, that the angle of the asymptotes becomes more and more obtuse, and that in the limit this special transformation changes into one in which the t' axis may have any upward direction whatsoever, while x' approaches more and more exactly to x. In view of this it is clear that group G_c in the limit when $c = \infty$, that is the group G_∞, becomes no other than that complete group which is appropriate to Newtonian mechanics. This being so, and since G_c is mathematically more intelligible than G_∞, it looks as though the thought might have struck some mathematician, fancy-free, that after all, as a matter of fact, natural phenomena do not possess an invariance with respect to the group G_∞, but rather with respect to a group G_c, with a certain finite c that is *extremely great* only in ordinary units of measure.

Eine solche Ahnung wäre ein außerordentlicher Triumph der reinen Mathematik gewesen. Nun, da die Mathematik hier nur mehr Treppenwitz bekundet, bleibt ihr doch die Genugtuung, daß sie dank ihren glücklichen Antezedenzien mit ihren in freier Fernsicht geschärften Sinnen die tiefgreifenden Konsequenzen einer solcher Ummodelung unserer Naturauffassung auf der Stelle zu erfassen vermag.

Ich will sogleich bemerken, um welchen Wert für c es sich schließlich handeln wird. Für c wird die *Fortpflanzungsgeschwindigkeit des Lichtes im leeren Raume* eintreten. Um weder vom Raum noch von Leere zu sprechen, können wir diese Größe wieder als das Verhältnis der elektrostatischen und der elektromagnetischen Einheit der Elektrizitätsmenge kennzeichnen.

Das Bestehen der Invarianz der Naturgesetze für die bezügliche Gruppe G_c würde nun so zu fassen sein:

Man kann aus der Gesamtheit der Naturerscheinungen durch sukzessiv gesteigerte Approximationen immer genauer ein Bezugsystem x, y, z und t, Raum und Zeit, ableiten, mittels dessen diese Erscheinungen sich dann nach bestimmten Gesetzen darstellen. Dieses Bezugsystem ist dabei aber durch die Erscheinungen keineswegs eindeutig festgelegt. *Man kann das Bezugsystem noch entsprechend den Transformationen der genannten Gruppe G_c beliebig verändern, ohne daß der Ausdruck der Naturgesetze sich dabei verändert.*

Z. B. kann man der beschriebenen Figur entsprechend auch t' Zeit benennen, muß dann aber im Zusammenhange damit notwendig den Raum durch die Mannigfaltigkeit der drei Parameter x', y, z definieren, wobei nun die physikalischen Gesetze mittels x', y, z, t' sich genau ebenso ausdrücken würden, wie mittels x, y, z, t. Hiernach würden wir dann in der Welt nicht mehr *den* Raum, sondern unendlich viele Räume haben, analog wie es im dreidimensionalen Räume unendlich viele Ebenen gibt. Die dreidimensionale Geometrie wird ein Kapitel der vierdimensionalen Physik. Sie erkennen, weshalb ich am Eingange sagte, Raum und Zeit sollen zu Schatten herabsinken und nur eine Welt an sich bestehen.

II.

Nun ist die Frage, welche Umstände zwingen uns die veränderte Auffassung von Raum und Zeit auf, widerspricht sie tatsächlich niemals den Erscheinungen, endlich gewährt sie Vorteile für die Beschreibung der Erscheinungen?

Bevor wir hierauf eingehen, sei eine wichtige Bemerkung vorangestellt. Haben wir Raum und Zeit irgendwie individualisiert, so entspricht einem ruhenden substantiellen Punkte als Weltlinie eine zur t-Achse parallele Gerade, einem gleichförmig bewegten substantiellen Punkte eine gegen die t-Achse geneigte Gerade, einem ungleichförmig bewegten substantiellen Punkte eine irgendwie gekrümmte Weltlinie. Fassen wir in einem beliebigen Weltpunkte x, y, z, t die dort durchlaufende Weltlinie auf, und finden wir sie dort parallel mit irgendeinem Radiusvektor OA' der vorhin genannten hyperboloidischen Schale, so können wir OA' als neue Zeitachse einführen, und bei den damit gegebenen neuen Begriffen von Raum und Zeit erscheint die Substanz in dem betreffenden Weltpunkte als ruhend. Wir wollen nun dieses fundamentale Axiom einführen:

Die in einem beliebigen Weltpunkte vorhandene Substanz kann stets bei geeigneter Festsetzung von Raum und Zeit als ruhend aufgefaßt werden.

Such a premonition would have been an extraordinary triumph for pure mathematics. Well, mathematics, though it now can display only staircase-wit, has the satisfaction of being wise after the event, and is able, thanks to its happy antecedents, with its senses sharpened by an unhampered outlook to far horizons, to grasp forthwith the far-reaching consequences of such a metamorphosis of our concept of nature.

I want to state at once what is the value of c with which we shall finally be dealing. It is the *velocity of the propagation of light in empty space*. To avoid speaking either of space or of emptiness, we can again define this magnitude as the ratio of the electromagnetic to the electrostatic unit of electricity.

The invariance of the laws of nature with respect to the relevant group G_c would have to be taken, then, in the following way:

From the totality of natural phenomena it is possible, by successively enhanced approximations, to derive more and more precisely a reference system x, y, z, t, space and time, by means of which these phenomena then present themselves in agreement with certain laws. But this reference system is by no means unequivocally determined by the phenomena. *It is still possible to change the reference system in accordance with the transformations of the group G_c, and still leave the expression of the laws of nature unaltered.*

For example, corresponding to the figure described above, we may also designate t' as time, but then must necessarily, connected to this, define space by the manifold x', y, z, in which case the physical laws would be expressed in exactly the same way by means of x', y, z, t' as by means of x, y, z, t. We would then have in the world no longer *the* 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. You see why I said at the outset that space and time are to fade away into shadows, and that only a world in itself will subsist.

II.

The question now is, what are the circumstances which force this changed conception of space and time upon us? Does it actually never contradict the phenomena? And finally, is it advantageous for describing the phenomena?

Before going into these questions, I must make an important remark. Once we have individualized space and time in some way, we have, as a world-line corresponding to a stationary substantial point, a straight line parallel to the t-axis; corresponding to a substantial point in uniform motion, a straight line at an angle to the axis of t; to a substantial point in non-uniform motion a somewhat curved world-line. If at any world-point x, y, z, t we take the world-line passing through that point, and find it parallel to any radius vector OA' of the above-mentioned hyperboloidal sheet, we can introduce OA' as a new axis of time, and with the new concepts of space and time thus given, the substance at the world-point concerned appears as at rest. We now want to introduce this fundamental axiom:

The substance existing at any world-point may always, with the appropriate fixation of space and time, be looked upon as at rest.

Das Axiom bedeutet, daß in jedem Weltpunkte stets der Ausdruck

$$c^2 dt^2 - dx^2 - dy^2 - dz^2$$

positiv ausfällt oder, was damit gleichbedeutend ist, daß jede Geschwindigkeit v stets kleiner als c ausfällt. Es würde danach für alle substantiellen Geschwindigkeiten c als obere Grenze bestehen und hierin eben die tiefere Bedeutung der Größe c liegen. In dieser anderen Fassung hat das Axiom beim ersten Eindruck etwas Mißfälliges. Es ist aber zu bedenken, daß nun eine modifizierte Mechanik Platz greifen wird, in der die Quadratwurzel aus jener Differentialverbindung zweiten Grades eingeht, so daß Fälle mit Überlichtgeschwindigkeit nur mehr eine Rolle spielen werden, etwa wie in der Geometrie Figuren mit imaginären Koordinaten.

Der *Anstoß* und wahre Beweggrund *für die Annahme der Gruppe*. G_c nun kam daher, daß die Differentialgleichung für die Fortpflanzung von Lichtwellen im leeren Raume jene Gruppe G_c besitzt.[2] Andererseits hat der Begriff starrer Körper nur in einer Mechanik mit der Gruppe G_∞ einen Sinn. Hat man nun eine Optik mit G_c, und gäbe es andererseits starre Körper, so ist leicht abzusehen, daß durch die zwei zu G_c und zu G_∞ gehörigen hyperboloidischen Schalen *eine* t-Richtung ausgezeichnet sein würde, und das würde weiter die Konsequenz haben, daß man an geeigneten starren optischen Instrumenten im Laboratorium einen Wechsel der Erscheinungen bei verschiedener Orientierung gegen die Fortschreitungsrichtung der Erde müßte wahrnehmen können. Alle auf dieses Ziel gerichteten Bemühungen, insbesondere ein berühmter Interferenzversuch von Michelson, hatten jedoch ein negatives Ergebnis. Um eine Erklärung hierfür zu gewinnen, bildete H. A. Lorentz eine Hypothese, deren Erfolg eben in der Invarianz der Optik für die Gruppe G_c liegt. Nach Lorentz soll jeder Körper, der eine Bewegung besitzt, in Richtung der Bewegung eine Verkürzung erfahren haben, und zwar bei einer Geschwindigkeit v im Verhältnisse

$$1 : \sqrt{1 - \frac{v^2}{c^2}}.$$

Diese Hypothese klingt äußerst phantastisch. Denn die Kontraktion ist nicht etwa als Folge von Widerständen im Äther zu denken, sondern rein als Geschenk von oben, als Begleitumstand des Umstandes der Bewegung.

Ich will nun an unserer Figur zeigen, daß die Lorentzsche Hypothese völlig äquivalent ist mit der neuen Auffassung von Raum und Zeit, wodurch sie viel verständlicher wird. Abstrahieren wir der Einfachheit wegen von y und z und denken uns eine räumlich eindimensionale Welt, so sind ein wie die t-Achse aufrechter und ein gegen die t-Achse geneigter Parallelstreifen (siehe Fig. 1) Bilder für den Verlauf eines ruhenden, bezüglich eines gleichförmig bewegten Körpers, der jedesmal eine konstante räumliche Ausdehnung behält.

---

[2] Eine wesentliche Anwendung dieser Tatsache findet sich bereits bei W. Voigt, Göttinger Nachr. 1887, S. 41.

The axiom signifies that at every world-point the expression

$$c^2dt^2 - dx^2 - dy^2 - dz^2$$

always has a positive value, or, what comes to the same, that any velocity v always proves less than c. Accordingly c would stand as the upper limit for all substantial velocities, and that is precisely what would reveal the deeper significance of the magnitude c. In this alternative form the first impression made by the axiom is not altogether pleasing. But we must bear in mind that a modified form of mechanics, in which the square root of this second order differential expression appears, will now make its way, so that cases with a velocity greater than that of light will henceforward play only some such part as that of figures with imaginary coordinates in geometry.

Now the *impulse* and true motivation for *assuming the group G_c* came from the fact that the differential equation for the propagation of light waves in empty space possesses the group G_c.[2] On the other hand, the concept of rigid bodies has meaning only in a mechanics satisfying the group G_∞. If we have a theory of optics with G_c, and if on the other hand there were rigid bodies, it is easy to see that *one* t-direction would be distinguished by the two hyperboloidal sheets appropriate to G_c and G_∞, and this would have the further consequence, that we should be able, by employing suitable rigid optical instruments in the laboratory, to perceive some alteration in the phenomena when the orientation with respect to the earth's motion is changed. But all efforts directed towards this goal, in particular a famous interference experiment of Michelson, had a negative result. To explain this failure, H.A. Lorentz set up a hypothesis, the success of which lies exactly in the invariance of optics with respect to the group G_c. According to Lorentz any moving body must have undergone a contraction in the direction of its motion. In particular, if the body has the velocity v, the contraction will be of the ratio

$$1 : \sqrt{1 - \frac{v^2}{c^2}}.$$

This hypothesis sounds extremely fantastical. For the contraction is not to be thought of as a consequence of resistance in the ether, but simply as a gift from above, as an accompanying circumstance of the circumstance of motion.

I now want to show by our figure that the Lorentzian hypothesis is completely equivalent to the new conception of space and time, which, indeed, makes the hypothesis much more intelligible. If for simplicity we disregard y and z, and imagine a world of one spatial dimension, then a parallel band, upright like the t-axis, and another inclining to the t-axis (see Fig. 1), are images for the career of a body at rest with respect to a uniformly moving body, which in each case preserves a constant spatial extent.

[2] An essential application of this fact can already be found in W. Voigt, Göttinger Nachrichten, 1887, p. 41.

Ist OA' parallel dem zweiten Streifen, so können wir t' als Zeit und x' als Raumkoordinate einführen, und es erscheint dann der zweite Körper als ruhend, der erste als gleichförmig bewegt. Wir nehmen nun an, daß der erste Körper als ruhend aufgefaßt die Länge l hat, d. h. der Querschnitt PP des ersten Streifens auf der x-Achse $= l \cdot OC$ ist, wo OC den Einheitsmaßstab auf der x-Achse bedeutet, und daß andererseits der zweite Körper *als ruhend aufgefaßt* die gleiche Länge l hat; letzteres heißt dann, daß der *parallel der x'-Achse* gemessene Querschnitt des zweiten Streifens, $Q'Q' = l \cdot OC'$ ist. Wir haben nunmehr in diesen zwei Körpern Bilder von zwei *gleichen* Lorentzschen Elektronen, einem ruhenden und einem gleichförmig bewegten. Halten wir aber an den ursprünglichen Koordinaten x, t fest, so ist als Ausdehnung des zweiten Elektrons der Querschnitt QQ seines zugehörigen Streifens *parallel der x-Achse* anzugeben. Nun ist offenbar, da $Q'Q' = l \cdot OC'$ ist, $QQ = l \cdot OD'$. Eine leichte Rechnung ergibt, wenn dx/dt für den zweiten Streifen $= v$ ist, $OD' = OC \cdot \sqrt{1 - \frac{v^2}{c^2}}$, also auch $PP : QQ = 1 : \sqrt{1 - \frac{v^2}{c^2}}$. Dies ist aber der Sinn der Lorentzschen Hypothese von der Kontraktion der Elektronen bei Bewegung. Fassen wir andererseits das zweite Elektron als ruhend auf, adoptieren also das Bezugsystem x', t', so ist als Länge des ersten der Querschnitt $P'P'$ seines Streifens parallel OC' zu bezeichnen, und wir würden in genau dem nämlichen Verhältnisse das erste Elektron gegen das zweite verkürzt finden; denn es ist in der Figur

$$P'P' : Q'Q' = OD : OC' = OD' : OC = QQ : PP.$$

Lorentz nannte die Verbindung t' von x und t *Ortszeit* des gleichförmig bewegten Elektrons und verwandte eine physikalische Konstruktion dieses Begriffs zum besseren Verständnis der Kontraktionshypothese. Jedoch scharf erkannt zu haben, daß die Zeit des einen Elektrons ebenso gut wie die des anderen ist, d. h. daß t und t' gleich zu behandeln sind, ist erst das Verdienst von A. Einstein.[3] Damit war nun zunächst die Zeit als ein durch die Erscheinungen eindeutig festgelegter Begriff abgesetzt. An dem Begriffe des Raumes rüttelten weder Einstein noch Lorentz, vielleicht deshalb nicht, weil bei der genannten speziellen Transformation, wo die x', t'-Ebene sich mit der x, t-Ebene deckt, eine Deutung möglich ist, als sei die x-Achse des Raumes in ihrer Lage erhalten geblieben. Über den Begriff des Raumes in entsprechender Weise hinwegzuschreiten, ist auch wohl nur als Verwegenheit mathematischer Kultur einzutaxieren. Nach diesem zum wahren Verständnis der Gruppe G_c jedoch unerläßlichen weiteren Schritt aber scheint mir das Wort *Relativitätspostulat* für die Forderung einer Invarianz bei der Gruppe G_c sehr matt. Indem der Sinn des Postulats wird, daß durch die Erscheinungen nur die in Raum und Zeit vierdimensionale Welt gegeben ist, aber die Projektion in Raum und in Zeit noch mit einer gewissen Freiheit vorgenommen werden kann, möchte ich dieser Behauptung eher den Namen *Postulat der absoluten Welt* (oder kurz Weltpostulat) geben.

[3] A. Einstein, Annalen der Physik **17** (1905), S. 891; Jahrbuch der Radioaktivität und Elektronik **4** (1907), S. 411.

If OA' is parallel to the second band, we can introduce t' as time and x' as space coordinate, and then the second body appears as at rest, the first as in uniform motion. We now assume that the first body has the length l when envisaged as at rest, that is, the cross section PP of the first band on the x-axis is equal to $l \cdot OC$, where OC denotes the unit of measure on the x-axis. On the other hand, we assume that the second body has the same length l *when envisaged as at rest*. The latter means that the cross section of the second band, *measured parallel to the x'-axis*, is $Q'Q' = l \cdot OC'$. We now have in these two bodies images of two *equal* Lorentzian electrons, one at rest and one in uniform motion. But if we retain the original coordinates x, t, we must give as the extent of the second electron the cross section QQ of its appropriate band *parallel to the x-axis*. Now since $Q'Q' = l \cdot OC'$, it is evident that $QQ = l \cdot OD'$. If $\mathrm{d}x/\mathrm{d}t$ for the second band is equal to v, an easy calculation gives $OD' = OC \cdot \sqrt{1 - \frac{v^2}{c^2}}$, and therefore also $PP : QQ = 1 : \sqrt{1 - \frac{v^2}{c^2}}$. But this is the meaning of Lorentz's hypothesis of the contraction of electrons in motion. If on the other hand we consider the second electron as at rest, and therefore adopt the reference system $x't'$, then the length of the first must be denoted by the cross section $P'P'$ of its band parallel to OC', and we would find the first electron in comparison with the second to be contracted in exactly the same proportion; for in the figure

$$P'P' : Q'Q' = OD : OC' = OD' : OC = QQ : PP.$$

Lorentz called the combination t' of x and t *local time*† of the electron in uniform motion, and used a physical construction of this concept, for the better understanding of the contraction hypothesis. But the credit of first recognizing clearly that the time of one of the electrons is just as good as that of the other, that is to say, that t and t' are to be treated identically, belongs to A. Einstein.[3] With this, time as a concept unequivocally determined by the phenomena was deposed from its high seat. Neither Einstein nor Lorentz made an attack on the concept of space, perhaps because in the above-mentioned special transformation, where the plane of x', t' coincides with the plane x, t, an interpretation is possible as if the x-axis of space had maintained its location. Indeed, stepping over the concept of space in such a way is something that we may see as only due to the audacity of mathematical culture. Nevertheless, this further step is indispensable for the true understanding of the group G_c, and when it has been taken, the word *relativity-postulate* for the requirement of an invariance with respect to the group G_c seems to me very feeble. Since the postulate comes to mean that only the four-dimensional world in space and time is given by the phenomena, whereas the projection in space and in time may still be undertaken with a certain degree of freedom, I prefer to call it the *postulate of the absolute world* (or briefly, the world-postulate).

[3] A. Einstein, Annalen der Physik **17** (1905), p. 891; Jahrbuch der Radioaktivität und Elektronik **4** (1907), p. 411.

III.

Durch das Weltpostulat wird eine gleichartige Behandlung der vier Bestimmungsstücke x, y, z, t möglich. Dadurch gewinnen, wie ich jetzt ausführen will, die Formen, unter denen die physikalischen Gesetze sich abspielen, an Verständlichkeit. Vor allem erlangt der Begriff der *Beschleunigung* ein scharf hervortretendes Gepräge.

Ich werde mich einer geometrischen Ausdrucksweise bedienen, die sich sofort darbietet, indem man im Tripel x, y, z stillschweigend von z abstrahiert. Einen beliebigen Weltpunkt O denke ich zum Raum-Zeit-Nullpunkt gemacht. Der *Kegel*

$$c^2t^2 - x^2 - y^2 - z^2 = 0$$

mit O als Spitze (Fig. 2) besteht aus zwei Teilen, einem mit Werten $t < 0$, einem anderen mit Werten $t > 0$. Der erste, der *Vorkegel von O*, besteht, sagen wir, aus allen Weltpunkten, die „Licht nach O senden", der zweite, der *Nachkegel von O*, aus allen Weltpunkten, die „Licht von O empfangen". Das vom Vorkegel allein begrenzte Gebiet mag *diesseits von O*, das vom Nachkegel allein begrenzte *jenseits von O* heißen. Jenseits O fällt die schon betrachtete hyperboloidische Schale

$$F = c^2t^2 - x^2 - y^2 - z^2 = 1, t > 0.$$

Das Gebiet *zwischen den Kegeln* wird erfüllt von den einschaligen hyperboloidischen Gebilden

$$-F = x^2 + y^2 + z^2 - c^2t^2 = k^2$$

zu allen konstanten positiven Werten k^2. Wichtig sind für uns die Hyperbeln mit O als Mittelpunkt, die auf den letzteren Gebilden liegen. Die einzelnen Äste dieser Hyperbeln mögen kurz die *Zwischenhyperbeln zum Zentrum O* heißen. Ein solcher Hyperbelast würde, als Weltlinie eines substantiellen Punktes gedacht, eine Bewegung repräsentieren, die für $t = -\infty$ und $t = +\infty$ asymptotisch auf die Lichtgeschwindigkeit c ansteigt.

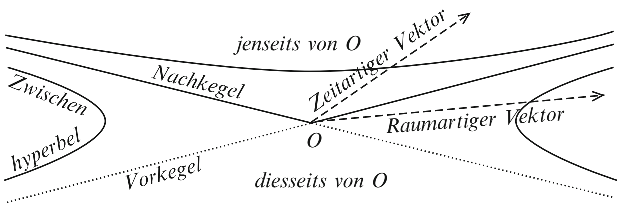


Fig. 2

III.

The world-postulate allows identical treatment of the four coordinates x, y, z, t. By this means, as I shall now show, the forms in which the laws of physics are displayed gain in intelligibility. In particular the concept of *acceleration* acquires a clear-cut character.

I will use a geometrical manner of expression, which suggests itself at once if we tacitly disregard z in the triple x, y, z. I take an arbitrary world-point O as the point of origin of space–time. The *cone*

$$c^2t^2 - x^2 - y^2 - z^2 = 0$$

with apex O (Fig. 2) consists of two parts, one with values $t < 0$, the other with values $t > 0$. The former, the *backward-cone of* O, consists, let us say, of all the world-points which "send light to O", the latter, the *forward-cone of* O, of all the world-points which "receive light from O". The territory bounded by the backward-cone alone we may call *before* O, that which is bounded by the forward-cone alone, *beyond* O‡. The hyperboloidal sheet already discussed

$$F = c^2t^2 - x^2 - y^2 - z^2 = 1, t > 0$$

lies after O. The territory *between the cones* is filled by the one-sheeted hyperboloidal figures

$$-F = x^2 + y^2 + z^2 - c^2t^2 = k^2$$

for all constant positive values of k^2. We are especially interested in the hyperbolas with O as centre, lying on the latter figures. The single branches of these hyperbolas may be called the *internal hyperbolas with centre* O. One of these branches, thought of as the world-line of a substantial point, would represent a motion which for $t = -\infty$ and $t = +\infty$ rises asymptotically to the velocity of light c.

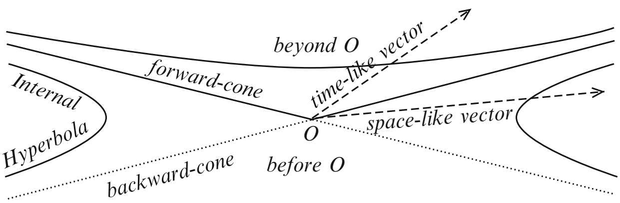


Fig. 2

Nennen wir in Analogie zum Vektorbegriff im Raume jetzt eine gerichtete Strecke in der Mannigfaltigkeit der x, y, z, t einen *Vektor*, so haben wir zu unterscheiden zwischen den *zeitartigen* Vektoren mit Richtungen von O nach der Schale $+F = 1, t > 0$ und den *raumartigen* Vektoren mit Richtungen von O nach $-F = 1$. Die Zeitachse kann jedem Vektor der ersten Art parallel laufen. Ein jeder Weltpunkt zwischen Vorkegel und Nachkegel von O kann durch das Bezugsystem als *gleichzeitig* mit O, aber ebensogut auch als *früher* als O oder als *später* als O eingerichtet werden. Jeder Weltpunkt diesseits O, ist notwendig stets früher, jeder Weltpunkt jenseits O notwendig stets später als O. Dem Grenzübergang zu $c = \infty$ würde ein völliges Zusammenklappen des keilförmigen Einschnittes zwischen den Kegeln in die ebene Mannigfaltigkeit $t = 0$ entsprechen. In den gezeichneten Figuren ist dieser Einschnitt absichtlich mit verschiedener Breite angelegt.

Einen beliebigen Vektor wie von O nach x, y, z, t zerlegen wir in die vier *Komponenten* x, y, z, t. Sind die Richtungen zweier Vektoren beziehungsweise die eines Radiusvektors OR von O an eine der Flächen $\mp F = 1$ und dazu einer Tangente RS im Punkte R der betreffenden Fläche, so sollen die Vektoren *normal* zueinander heißen. Danach ist

$$c^2 t t_1 - x x_1 - y y_1 - z z_1 = 0$$

die Bedingung dafür, daß die Vektoren mit den Komponenten x, y, z, t und x_1, y_1, z_1, t_1 normal zueinander sind.

Für die *Beträge* von Vektoren der verschiedenen Richtungen sollen die *Einheitsmaßstäbe* dadurch fixiert sein, daß einem raumartigen Vektor von O nach $-F = 1$ stets der Betrag 1, einem zeitartigen Vektor von O nach $+F = 1, t > 0$ stets der Betrag $1/c$ zugeschrieben wird.

Denken wir uns nun in einem Weltpunkte $P(x, y, z, t)$ die dort durchlaufende Weltlinie eines substantiellen Punktes, so entspricht danach dem zeitartigen Vektorelement dx, dy, dz, dt im Fortgang der Linie der Betrag

$$d\tau = \frac{1}{c}\sqrt{c^2 dt^2 - dx^2 - dy^2 - dz^2}.$$

Das Integral $\int d\tau = \tau$ dieses Betrages auf der Weltlinie von irgendeinem fixierten Ausgangspunkte P_0 bis zu dem variablen Endpunkte P geführt, nennen wir die *Eigenzeit* des substantiellen Punktes in P. Auf der Weltlinie betrachten wir x, y, z, t, d. s. die Komponenten des Vektors OP, als Funktionen der Eigenzeit τ, bezeichnen deren erste Differentialquotienten nach τ mit $\dot{x}, \dot{y}, \dot{z}, \dot{t}$, deren zweite Differentialquotienten nach τ mit $\ddot{x}, \ddot{y}, \ddot{z}, \ddot{t}$ und nennen die zugehörigen Vektoren, die Ableitung des Vektors OP nach τ den *Bewegungsvektor in* P und die Ableitung dieses Bewegungsvektors nach τ den *Beschleunigungsvektor in* P. Dabei gilt

$$c^2\dot{t}^2 - \dot{x}^2 - \dot{y}^2 - \dot{z}^2 = c^2,$$
$$c^2\dot{t}\ddot{t} - \dot{x}\ddot{x} - \dot{y}\ddot{y} - \dot{z}\ddot{z} = 0,$$

If we now, on the analogy of vectors in space, call a directed length in the manifold x, y, z, t a *vector*, then we have to distinguish between the *time-like* vectors with directions from O to the sheet $+F = 1, t > 0$, and the *space-like* vectors with directions from O to $-F = 1$. The time axis may run parallel to any vector of the first kind. Any world-point between the backward-cone and the forward-cone of O can be arranged by means of a reference system so as to be *simultaneous* with O, but also just as well so as to be *earlier* than O or *later* than O. Any world-point within the backward-cone of O is necessarily always before O; any world-point within the forward-cone is necessarily always after O. Corresponding to passing to the limit $c = \infty$ we would have a complete collapse of the wedge-shaped segment between the cones into the flat manifold $t = 0$. In the figures this segment is intentionally drawn with different widths.

We divide up any vector we choose, e.g. that from O to x, y, z, t, into the four *components* x, y, z, t. If the directions of two vectors are, respectively, that of a radius vector OR from O to one of the surfaces $\mp F = 1$, and that of a tangent RS at the point R of the same surface, the vectors are said to be *normal* to one another. Thus the condition that the vectors with components x, y, z, t and x_1, y_1, z_1, t_1 are normal to each other is

$$c^2tt_1 - xx_1 - yy_1 - zz_1 = 0.$$

For the *magnitudes* of vectors in different directions the *units of measure* are to be fixed by assigning to a space-like vector from O to $-F = 1$ always the magnitude 1, and to a time-like vector from O to $+F = 1, t > 0$ always the magnitude $1/c$.

If we imagine at a world-point $P(x, y, z, t)$ the world-line of a substantial point running through that point, the magnitude corresponding to the time-like vector dx, dy, dz, dt laid off along the line is therefore

$$d\tau = \frac{1}{c}\sqrt{c^2dt^2 - dx^2 - dy^2 - dz^2}.$$

The integral $\int d\tau = \tau$ of this magnitude, taken along the world-line from any fixed starting-point P_0 to the variable endpoint P, we call the *proper time* of the substantial point at P[§]. On the world-line we regard x, y, z, t – the components of the vector OP – as functions of the proper time τ; denote their first differential coefficients with respect to τ by $\dot{x}, \dot{y}, \dot{z}, \dot{t}$; their second differential coefficients with respect to τ by $\ddot{x}, \ddot{y}, \ddot{z}, \ddot{t}$; and give names to the corresponding vectors, calling the derivative of the vector OP the *velocity vector at* P and the derivative of the velocity vector with respect to τ the *acceleration vector at* P. Hence, since

$$c^2\dot{t}^2 - \dot{x}^2 - \dot{y}^2 - \dot{z}^2 = c^2,$$

we have

$$c^2\dot{t}\ddot{t} - \dot{x}\ddot{x} - \dot{y}\ddot{y} - \dot{z}\ddot{z} = 0,$$

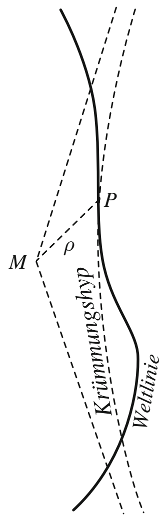


Fig. 3

d. h. der Bewegungsvektor ist der zeitartige Vektor in Richtung der Weltlinie in P vom Betrage 1, und der Beschleunigungsvektor in P ist normal zum Bewegungsvektor in P, also jedenfalls ein raumartiger Vektor.

Nun gibt es, wie man leicht einsieht, einen bestimmten Hyperbelast, der mit der Weltlinie in P drei unendlich benachbarte Punkte gemein hat, und dessen Asymptoten Erzeugende eines Vorkegels und eines Nachkegels sind (siehe unten Fig. 3). Dieser Hyperbelast heiße die *Krümmungshyperbel* in P. Ist M das Zentrum dieser Hyperbel, so handelt es sich also hier um eine Zwischenhyperbel zum Zentrum M. Es sei ρ der Betrag des Vektors MP, *so erkennen wir den Beschleunigungsvektor in P als den Vektor in Richtung MP vom Betrage c^2/ρ.*

Sind $\ddot{x}, \ddot{y}, \ddot{z}, \ddot{t}$ sämtlich Null, so reduziert sich die Krümmungshyperbel auf die in P die Weltlinie berührende Gerade, und es ist $\rho = \infty$ zu setzen.

IV.

Um darzutun, daß die Annahme der Gruppe G_c für die physikalischen Gesetze nirgends zu einem Widerspruche führt, ist es unumgänglich, eine Revision der gesamten Physik auf Grund der Voraussetzung dieser Gruppe vorzunehmen. Diese Revision ist bereits in einem gewissen Umfange erfolgreich geleistet für Fragen der Thermodynamik und Wärmestrahlung [4], für die elektromagnetischen Vorgänge, endlich für die Mechanik unter Aufrechterhaltung des Massenbegriffs.[5]

[4] M. Planck, “Zur Dynamik bewegter Systeme”, Berliner Berichte 1907, S. 542. (auch Annalen der Physik **26** (1908), S. 1.)

[5] H. Minkowski, “Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körpern”, Göttinger Nachrichten 1908, S. 53.

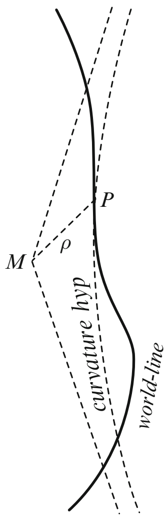


Fig. 3

i.e. the velocity vector is the time-like vector of unit magnitude in the direction of the world-line at P. It also follows that the acceleration vector at P is normal to the velocity vector at P, and is therefore in any case a space-like vector.

Now, as is readily seen, there is a definite hyperbola-branch, which has three infinitely close points in common with the world-line at P, and whose asymptotes are generator of a backward-cone and a forward-cone. (Fig. 3). Let this hyperbola be called the *curvature hyperbola* at P. If M is the centre of this hyperbola, we here have to do with an internal hyperbola with centre M. Let ρ be the magnitude of the vector MP; *then we recognize the acceleration vector at P as the vector in the direction MP of magnitude c^2/ρ.*

If $\ddot{x}, \ddot{y}, \ddot{z}, \ddot{t}$ are all zero, the curvature hyperbola reduces to the straight line touching the world-line in P, and we must put $\rho = \infty$.

IV.

To substantiate that the assumption of the group G_c for the laws of physics never leads to a contradiction, it is unavoidable to undertake a revision of the whole of physics on the basis of this assumption. This revision has to some extent already been successfully carried out for questions of thermodynamics and heat radiation[4], for the electromagnetic processes, and finally, with the retention of the concept of mass, for mechanics.[5]

---

[4] M. Planck, "Zur Dynamik bewegter Systeme", Berliner Berichte 1907, p. 542. (Also in Annalen der Physik **26** (1908), p. 1.)

[5] H. Minkowski, "Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körpern", Göttinger Nachrichten 1908, p. 53.

Für letzteres Gebiet ist vor allem die Frage aufzuwerfen: Wenn eine Kraft mit den Komponenten X, Y, Z nach den Raumachsen in einem Weltpunkte $P(x, y, z, t)$ angreift, wo der Bewegungsvektor $\dot{x}, \dot{y}, \dot{z}, \dot{t}$ ist, als welche Kraft ist diese Kraft bei einer beliebigen Änderung des Bezugsystemes aufzufassen? Nun existieren gewisse erprobte Ansätze über die ponderomotorische Kraft im elektromagnetischen Felde in den Fällen, wo die Gruppe G_c unzweifelhaft zuzulassen ist. Diese Ansätze führen zu der einfachen Regel: *Bei Änderung des Bezugsystemes ist die vorausgesetzte Kraft derart als Kraft in den neuen Raumkoordinaten anzusetzen, daß dabei der zugehörige Vektor mit den Komponenten*

$$\dot{t}X,\ \dot{t}Y,\ \dot{t}Z,\ \dot{t}T,$$

wo

$$T = \frac{1}{c^2}\left(\frac{\dot{x}}{\dot{t}}X + \frac{\dot{y}}{\dot{t}}Y + \frac{\dot{z}}{\dot{t}}Z\right)$$

die durch c^2 dividierte Arbeitsleistung der Kraft im Weltpunkte ist, sich unverändert erhält. Dieser Vektor ist stets normal zum Bewegungsvektor in P. Ein solcher, zu einer Kraft in P gehörender Kraftvektor soll ein *bewegender Kraftvektor in P* heißen.

Nun werde die durch P laufende Weltlinie von einem substantiellen Punkte mit konstanter *mechanischer Masse m* beschrieben. Das m-fache des Bewegungsvektors in P heiße der *Impulsvektor in P*, das m-fache des Beschleunigungsvektors in P der *Kraftvektor der Bewegung in P*. Nach diesen Definitionen lautet das Gesetz dafür, wie die Bewegung eines Massenpunktes bei gegebenem bewegenden Kraftvektor statthat:[6]

Der Kraftvektor der Bewegung ist gleich dem bewegenden Kraftvektor.

Diese Aussage faßt vier Gleichungen für die Komponenten nach den vier Achsen zusammen, wobei die vierte, weil von vornherein beide genannten Vektoren normal zum Bewegungsvektor sind, sich als eine Folge der drei ersten ansehen läßt. Nach der obigen Bedeutung von T stellt die vierte zweifellos den Energiesatz dar. Als *kinetische Energie* des Massenpunktes ist daher das *c^2-fache der Komponente des Impulsvektors nach der t-Achse* zu definieren. Der Ausdruck hierfür ist

$$mc^2\frac{dt}{d\tau} = mc^2 \Big/ \sqrt{1 - \frac{v^2}{c^2}},$$

d. i. nach Abzug der additiven Konstante mc^2 der Ausdruck $\frac{1}{2}mv^2$ der Newtonschen Mechanik bis auf Größen von der Ordnung $1/c^2$ Sehr anschaulich erscheint hierbei die *Abhängigkeit der Energie vom Bezugsysteme*. Da nun aber die t-Achse in die Richtung jedes zeitartigen Vektors gelegt werden kann, so enthält andererseits der Energiesatz, für jedes mögliche Bezugsystem gebildet, bereits das ganze System der Bewegungsgleichungen.

[6] H. Minkowski, a. a. O. S. 107. Vgl. auch M. Planck, Verhandlungen der physikalischen Gessellschaft **4** (1906), S. 136.

For this last branch of physics it is of prime importance to raise the following question: When a force with the spatial components X, Y, Z acts at a world-point $P(x, y, z, t)$, where the velocity vector is $\dot{x}, \dot{y}, \dot{z}, \dot{t}$, what must we take this force to be if the reference system is changed arbitrarily? Now there exist certain tested approaches as to the ponderomotive force in the electromagnetic field, where the group G_c is undoubtedly admissible. These approaches lead up to the following simple rule: *When the reference system is changed, the force in question transforms into a force in the new spatial coordinates. This happens in such a way that the corresponding vector with the components*

$$\dot{t}X, \dot{t}Y, \dot{t}Z, \dot{t}T$$

where

$$T = \frac{1}{c^2}\left(\frac{\dot{x}}{\dot{t}}X + \frac{\dot{y}}{\dot{t}}Y + \frac{\dot{z}}{\dot{t}}Z\right)$$

is the rate at which work is done by the force at the world-point divided by c^2, remains unchanged. This vector is always normal to the velocity vector at P. Such a force vector, belonging to a force at P, is to be called a *motive force vector* at P.

Let us now assume that the world-line passing through P corresponds to a substantial point with constant *mechanical mass m*. Let the velocity vector at P multiplied by m be called the *momentum vector at P*, and the acceleration vector at P multiplied by m, be called the *force vector of the motion at P*. With these definitions, the law of motion for a mass point with given force vector is:[6]

The Force Vector of the Motion is Equal to the Motive Force Vector.

This assertion comprises four equations for the components corresponding to the four axes, and since both vectors mentioned are normal to the velocity vector from the very outset, the fourth equation may be looked upon as a consequence of the other three. In accordance with the above meaning of T, the fourth equation undoubtedly represents the law of energy conservation. Thus, the *kinetic energy* of the mass point is defined as the *component of the momentum vector along the t-axis multiplied by c^2*. The expression for this is

$$mc^2\frac{dt}{d\tau} = mc^2 \Big/ \sqrt{1 - \frac{v^2}{c^2}},$$

which is, after removal of the additive constant mc^2, the expression $\frac{1}{2}mv^2$ of Newtonian mechanics down to magnitudes of the order $1/c^2$. It comes out very clearly in this way, how *the energy depends on the reference system*. On the other hand, since the t-axis may be laid in the direction of any time-like vector, the law of energy conservation, built for all possible reference systems, already contains the whole system of the equations of motion.

[6] H. Minkowski, loc. cit., p. 107. CF also M. Planck, Verhandlungen der physikalischen Gessellschaft **4** (1906), p. 136.

Diese Tatsache behält bei dem erörterten Grenzübergang zu $c = \infty$ ihre Bedeutung auch für den axiomatischen Aufbau der Newtonschen Mechanik und ist in solchem Sinne hier bereits von Herrn J. R. Schütz[7] wahrgenommen worden.

Man kann von vornherein das Verhältnis von Längeneinheit und Zeiteinheit derart wählen, daß die natürliche Geschwindigkeitsschranke $c = 1$ wird. Führt man dann noch $\sqrt{-1} \cdot t = s$ an Stelle von t ein, so wird der quadratische Differentialausdruck

$$d\tau^2 = -dx^2 - dy^2 - dz^2 - ds^2,$$

also völlig symmetrisch in x, y, z, s, und diese Symmetrie überträgt sich auf ein jedes Gesetz, das dem Weltpostulate nicht widerspricht. Man kann danach das Wesen dieses Postulates mathematisch sehr prägnant in die mystische Formel kleiden:

$$3.10^5 \text{ Km} = \sqrt{-1}\text{sek.}$$

V.

Die durch das Weltpostulat geschaffenen Vorteile werden vielleicht durch nichts so schlagend belegt wie durch Angabe der von einer *beliebig bewegten punktförmigen Ladung* nach der Maxwell-Lorentzschen Theorie ausgehenden Wirkungen. Denken wir uns die Weltlinie eines solchen punktförmigen Elektrons mit der Ladung e und führen auf ihr die Eigenzeit τ ein von irgendeinem Anfangspunkte aus. Um das vom Elektron in einem beliebigen Weltpunkte P_1 veranlaßte Feld zu haben, konstruieren wir den zu P_1 gehörigen Vorkegel (Fig. 4). Dieser trifft die unbegrenzte Weltlinie des Elektrons, weil deren Richtungen überall die von zeitartigen Vektoren sind, offenbar in einem einzigen Punkte P. Wir legen in P an die Weltlinie die Tangente und konstruieren durch P_1 die Normale $P_1 Q$ auf diese Tangente. Der Betrag von $P_1 Q$ sei r. Als der Betrag von PQ ist dann gemäß der Definition eines Vorkegels r/c zu rechnen. *Nun stellt der Vektor in Richtung PQ vom Betrage e/r in seinen Komponenten nach den x-, y-, z-Achsen das mit c multiplizierte Vektorpotential, in der Komponente nach der t-Achse das skalare Potential des von e erregten Feldes für den Weltpunkt P_1 vor.* Hierin liegen die von A. Liénard und von E. Wiechert aufgestellten Elementargesetze.[8]

Bei der Beschreibung des vom Elektron hervorgerufenen Feldes selbst tritt sodann hervor, daß die Scheidung des Feldes in elektrische und magnetische Kraft eine relative ist mit Rücksicht auf die zugrunde gelegte Zeitachse; am übersichtlichsten sind beide Kräfte zusammen zu beschreiben in einer gewissen, wenn auch nicht völligen Analogie zu einer Kraftschraube der Mechanik.

---

[7] I.R. Schütz, "Das Prinzip der absoluten Erhaltung der Energie", Göttinger Nachrichten 1897, S. 110.

[8] A. Liénard, "Champ électrique et magnétique produit par une charge concentré en un point et animée d'un mouvement quelconque", L'Éclairage électrique **16** (1898), S. 5, 53, 106; E. Wiechert, "Elektrodynamische Elementargesetze", Arch. nèerl (2), **5**, (1900), S. 549.

At the limiting transition which we have discussed, to $c = \infty$, this fact retains its importance even for the axiomatic composition of Newtonian mechanics. As such, it has already been appreciated by I.R Schütz.[7]

We can, from the very start, decide upon the ratio of the units of length and time in such a way that the natural limit of velocity becomes $c = 1$. If we then introduce $\sqrt{-1} \cdot t = s$ in place of t, the quadratic differential expression

$$d\tau^2 = -dx^2 - dy^2 - dz^2 - ds^2$$

thus becomes perfectly symmetrical in x, y, z, s. This symmetry is communicated to any law which does not contradict the world-postulate. Thus the essence of this postulate can be clothed mathematically in a very pregnant manner in the mystic formula

$$3 \cdot 10^5 km = \sqrt{-1} secs.$$

V.

The advantages afforded by the world-postulate will perhaps be most strikingly exemplified by indicating the effects proceeding from a *point charge in any kind of motion* according to Maxwell–Lorentz theory. Let us imagine the world-line of such a point electron with the charge e, and introduce upon it the proper time τ from any initial point. In order to find the field caused by the electron at an arbitrary world-point P_1, we construct the backward-cone belonging to P_1 (Fig. 4). The cone evidently meets the world-line of the electron at a single point P, for the directions of the world-line are everywhere those of time-like vectors. We draw the tangent to the world-line at P, and construct through P_1 the normal P_1Q to this tangent. Let the magnitude of P_1Q be r. Then, by the definition of a backward-cone, the length of PQ must be r/c. *Now the vector in the direction PQ of magnitude e/r represents by its components along the x-, y-, z-axes, the vector potential multiplied by c, and by the component along the t-axis, the scalar potential of the field excited by e at the world-point* P_1. Herein lie the elementary laws formulated by A. Liénard and E. Wiechert.[8]

Then in the description of the field produced by the electron we see that the separation of the field into electric and magnetic force is a relative one with regard to the chosen time-axis; the most perspicuous way of describing the two forces is in a unified manner, in a certain analogy with the wrench in mechanics, even though the analogy is not a complete one.

[7] I.R. Schütz, "Das Prinzip der absoluten Erhaltung der Energie", Göttinger Nachrichten 1897, p. 110.

[8] A. Liénard, "Champ électrique et magnétique produit par une charge concentré en un point et animée d'un mouvement quelconque", L'Éclairage électrique **16** (1898), pp. 5, 53, 106; E. Wiechert, "Elektrodynamische Elementargesetze", Arch. nèerl (2), **5**, (1900), p. 549.

Ich will jetzt die *von einer beliebig bewegten punktförmigen Ladung auf eine andere beliebig bewegte punktförmige Ladung ausgeübte ponderomotorische Wirkung* beschreiben. Denken wir uns durch den Weltpunkt P_1 die Weltlinie eines zweiten punktförmigen Elektrons von der Ladung e_1 führend. Wir bestimmen P, Q, r wie vorhin, konstruieren sodann (Fig. 4) den Mittelpunkt der Krümmungshyperbel in P, endlich die Normale MN von M aus auf eine durch P parallel zu QP_1 gedachte Gerade.

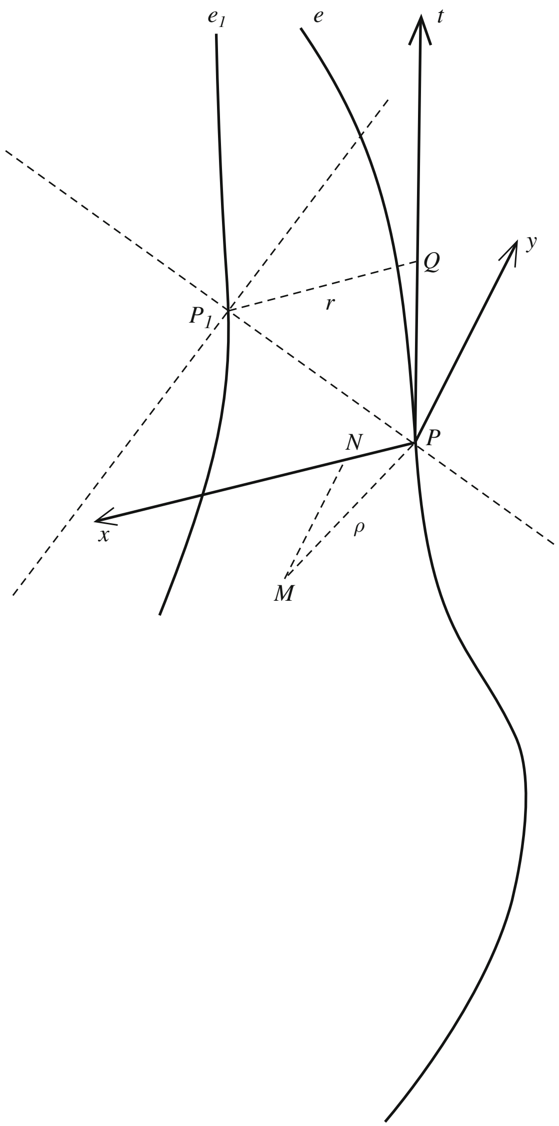


Fig. 4

I now want to describe *the ponderomotive effect of an arbitrarily moving point charge on another arbitrarily moving point charge*. Let us imagine the world-line of a second point electron of the charge e_1, passing through the world-point P_1. We define P, Q, r as before, then construct (Fig. 4) the centre M of the curvature hyperbola at P, and finally the normal MN from M to a straight line imagined through P parallel to QP_1.

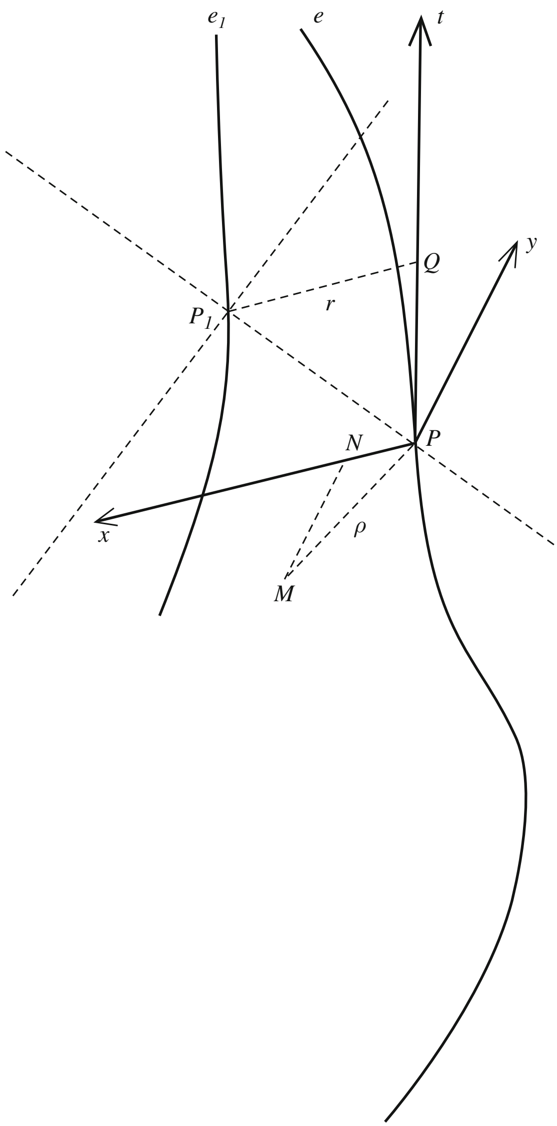


Fig. 4

Wir legen nun, mit P als Anfangspunkt, ein Bezugsystem folgendermaßen fest, die t-Achse in die Richtung PQ, die x-Achse in die Richtung QP_1 die y-Achse in die Richtung MN, womit schließlich auch die Richtung der z-Achse als normal zu den t-, x-, y-Achsen bestimmt ist. Der Beschleunigungsvektor in P sei $\ddot{x}, \ddot{y}, \ddot{z}, \ddot{t}$, der Bewegungsvektor in P_1 sei $\dot{x}_1, \dot{y}_1, \dot{z}_1, \dot{t}$, *Jetzt lautet der von dem ersten beliebig bewegten Elektron e auf das zweite beliebig bewegte Elektron e_1 in P_1 ausgeübte bewegende Kraftvektor:*

$$-e\,e_1 \left(\dot{t}_1 - \frac{\dot{x}_1}{c} \right) \mathfrak{R},$$

wobei für die Komponenten $\mathfrak{R}_x$, $\mathfrak{R}_y$, $\mathfrak{R}_z$, $\mathfrak{R}_t$ des Vektors $\mathfrak{R}$ die drei Relationen bestehen:

$$c\mathfrak{R}_t - \mathfrak{R}_x = \frac{1}{r^2}, \quad \mathfrak{R}_y = \frac{\ddot{y}}{c^2 r}, \quad \mathfrak{R}_z = 0$$

und viertens dieser Vektor $\mathfrak{R}$ normal zum Bewegungsvektor in P_1 ist und durch diesen Umstand allein in Abhängigkeit von dem letzteren Bewegungsvektor steht.

Vergleicht man mit dieser Aussage die bisherigen Formulierungen[9] des nämlichen Elementargesetzes über die ponderomotorische Wirkung bewegter punktförmiger Ladungen aufeinander, so wird man nicht umhin können, zuzugeben, daß die hier in Betracht kommenden Verhältnisse ihr inneres Wesen voller Einfachheit erst in vier Dimensionen enthüllen, auf einen von vornherein aufgezwungenen dreidimensionalen Raum aber nur eine sehr verwickelte Projektion werfen.

In der dem Weltpostulate gemäß reformierten Mechanik fallen die Disharmonien, die zwischen der Newtonschen Mechanik und der modernen Elektrodynamik gestört haben, von selbst aus. Ich will noch die Stellung des Newtonschen Attraktionsgesetzes zu diesem Postulate berühren. Ich will annnehmen, wenn zwei Massenpunkte m, m_1 ihre Weltlinien beschreiben, werde von m auf m_1 ein bewegender Kraftvektor ausgeübt genau von dem soeben im Falle von Elektronen angegebenen Ausdruck, nur daß statt $-e\,e_1$ jetzt $+mm_1$ treten soll. Wir betrachten nun speziell den Fall, daß der Beschleunigungsvektor von m konstant Null ist, wobei wir dann t so einführen mögen, daß m als ruhend aufzufassen ist, und es erfolge die Bewegung von m_1 allein mit jenem von m herrührenden bewegenden Kraftvektor. Modifizieren wir nun diesen angegebenen Vektor zunächst durch Hinzusetzen des Faktors $\dot{t}^{-1} = \sqrt{1 - \frac{v^2}{c^2}}$, der bis auf Größen von der Ordnung $1/c^2$ auf 1 hinauskommt, so zeigt sich[10], daß für die Orte x_1, y_1, z_1 von m_1 und ihren zeitlichen Verlauf genau wieder die Keplerschen Gesetze hervorgehen würden, nur daß dabei an Stelle der Zeiten t_1 die Eigenzeiten τ_1 von m_1 eintreten würden.

[9] K. Schwazschild, Göttinger Nachrichten 1903, S. 132; H.A. Lorentz, Enzyklopädie der mathematischen Wissenschaften V, Artikel 14, S. 199.

[10] H. Minkowski, a. a. O. S. 110.

With P as starting point we now choose a reference system as follows: The t-axis in the direction PQ, the x-axis in the direction QP_1, the y-axis in the direction MN. From this the direction of the z-axis is determined by the requirement of being normal to the t-, x-, y-axes. Let the acceleration vector at P be $\ddot{x}, \ddot{y}, \ddot{z}, \ddot{t}$, the velocity vector at P_1 be $\dot{x}, \dot{y}, \dot{z}, \dot{t}$. *The motive force vector exerted at P_1 by the first arbitrarily moving electron e on the second arbitrarily moving electron e_1 now takes the form*

$$-ee_1\left(\dot{t}_1 - \frac{\dot{x}_1}{c}\right)\mathfrak{R},$$

where the components $\mathfrak{R}_x, \mathfrak{R}_y, \mathfrak{R}_z, \mathfrak{R}_t$ of the vector $\mathfrak{R}$ satisfy the three relations:

$$c\mathfrak{R}_t - \mathfrak{R}_x = \frac{1}{r^2}, \quad \mathfrak{R}_y = \frac{\ddot{y}}{c^2 r}, \quad \mathfrak{R}_z = 0$$

and where, fourthly, this vector $\mathfrak{R}$ is normal to the velocity vector at P_1, and through this circumstance alone stands in dependence on the latter velocity vector.

When we compare this statement with previous formulations[9] of the same elementary law of the ponderomotive action of moving point charges on one another, we are compelled to admit that it is only in four dimensions that the relations here taken under consideration reveal their inner being in full simplicity, and that on a three dimensional space forced upon us from the very beginning they cast only a very tangled projection.

In mechanics as reformed in accordance with the world-postulate, the disturbing lack of harmony between Newtonian mechanics and modern electrodynamics disappears of its own accord. Before concluding I want to touch upon the attitude of Newton's law of attraction toward this postulate. I shall assume that when two mass points m, m_1 follow their world-lines, a motive force vector is exerted by m on m_1, of exactly the same form as that just given for the case of electrons, except that $+mm_1$ must now take the place of $-ee_1$. We now especially consider the case where the acceleration vector of m is constantly zero. Let us then introduce t in such a way that m is to be taken as at rest, and let only m_1 move under the motive force vector which proceeds from m. If we now modify this given vector in the first place by adding the factor $\dot{t}^{-1} = \sqrt{1 - \frac{v^2}{c^2}}$, which, up to the magnitudes of order $1/c^2$, is equal to 1, it can be seen[10] that for the positions x_1, y_1, z_1 of m_1 and their development in time, we should arrive exactly at Kepler's laws again, except that the proper times τ_1 of m_1 would take the place of the times t_1.

[9] K. Schwazschild, Göttinger Nachrichten 1903, p. 132; H.A. Lorentz, Enzyklopädie der mathematischen Wissenschaften V, Artikel 14, p. 199.

[10] H. Minkowski, loc. cit., p. 110.

Auf Grund dieser einfachen Bemerkung läßt sich dann einsehen, daß das vorgeschlagene Anziehungsgesetz verknüpft mit der neuen Mechanik nicht weniger gut geeignet ist die astronomischen Beobachtungen zu erklären als das Newtonsche Anziehungsgesetz verknüpft mit der Newtonschen Mechanik.

Auch die Grundgleichungen für die elektromagnetischen Vorgänge in ponderablen Körpern fügen sich durchaus dem Weltpostulate. Sogar die von Lorentz gelehrte Ableitung dieser Gleichungen auf Grund von Vorstellungen der Elektronentheorie braucht zu dem Ende keineswegs verlassen zu werden, wie ich anderwärts zeigen werde.

Die ausnahmslose Gültigkeit des Weltpostulates ist, so möchte ich glauben, der wahre Kern eines elektromagnetischen Weltbildes, der von Lorentz getroffen, von Einstein weiter herausgeschält, nachgerade vollends am Tage liegt. Bei der Fortbildung der mathematischen Konsequenzen werden genug Hinweise auf experimentelle Verifikationen des Postulates sich einfinden, um auch diejenigen, denen ein Aufgeben altgewohnter Anschauungen unsympathisch oder schmerzlich ist, durch den Gedanken an eine prästabilierte Harmonie zwischen der reinen Mathematik und der Physik auszusöhnen.

From this simple remark it may then be seen that the proposed law of attraction combined with the new mechanics is no less well suited to explain the astronomical observations than the Newtonian law of attraction combined with Newtonian mechanics.

The fundamental equations for electromagnetic processes in ponderable bodies also comply completely with the world-postulate. As I shall show elsewhere, it is not even necessary to abandon the derivation of these fundamental equations starting from ideas of electron theory, as taught by Lorentz, in order for them to comply with the world-postulate.

The validity without exception of the world-postulate, I like to think, is the true core of an electromagnetic world picture, which, hit upon by Lorentz, revealed further by Einstein, now lies open in the full light of day. In the development of its mathematical consequences there will be ample suggestions for experimental verifications of the postulate, which will suffice to conciliate even those to whom the abandonment of old-established views is unsympathetic or painful, by the thought of a pre-established harmony between pure mathematics and physics.

Notes of the Translator

D. Lehmkuhl

Oriel College, Oxford University and Institute of Philosophy and Interdisciplinary Centre for Science and Technology Studies, University of Wuppertal
e-mail: dennis.lehmkuhl@uni-wuppertal.de

The English version of Minkowski's text is due to a critical reassesment of W. Perret and G.B. Jeffrey translation of Minkowski's text, which has become standard as part of their translation of the collection entitled "The Principle of Relativity". The latter contained original papers by Lorentz, Einstein, Minkowski and Weyl, and was first published by Dover in 1952; it saw numerous reprints, the last one in 2000. I found that Perret and Jeffrey did a wonderful job, but that there was still plenty of room to improve their translation; in some cases their translation even fostered a misinterpretation of Minkowski. The most crucial changes as compared to their translation are summarised in the following notes.

[⋆]Arguably, this is one of the most famous sentences in the history of physics, and Perrett and Jeffery's translation captures the spirit beautifully; I do not want to change it. But I reckon that it might be fruitful for the reader to see a much more literal but less beautiful translation of Minkowski's words: "From this hour on space by itself and time by itself shall dwindle to shadows, and only a kind of union of the two shall preserve independence."

[†]This is the translation of Lorentz's *Ortszeit* that has become established in the English-speaking literature; a more literal translation would be '*placetime*'.

‡Minkowski's original terms for the backward-cone and the forward-cone are 'Vorkegel' and 'Nachkegel'. Perrett and Jeffery translated these terms as 'front cone' and 'back cone'. Even though this is quite a literal translation in a context where 'vor' and 'nach' refer to something *spatial*, this is not so if the terms are used in a temporal or causal context, as we have it here. In this case, Perrett and Jeffery's translation reverses the intuitions as compared to the original, for then the German word 'vor' corresponds to 'before', not 'front', while 'nach' refers to 'after', not 'back'. Even though Perrett and Jeffery clarify the situation in their translation of the following sentence, which I follow apart from writing 'beyond' rather than 'after', I decided to translate the names of the two parts of a light cone differently in order to preserve the intuitions triggered by Minkowski's original choice of words. (It happens that this translation is also more in accord with modern English terminology, where one often speaks of 'backward light cones' and 'forward light cones'.) With respect to the second sentence, it should be noted that even though 'before' and 'beyond' capture the content, the translation does not capture the full poetical spirit of Minkowski's original words: he writes 'diesseits' (used for 'now and in the future' but also for 'the world of the living') and 'jenseits' (used for 'beyond us' but also for 'the afterlife').

§The word originally chosen by Minkowski is "Eigenzeit". The translation as 'proper time' is long established, but it is interesting to note that 'eigentime' would be a more literal translation, corresponding to 'eigenvectors' in vector algebra.

Part I
The Impact of Minkowski Spacetime on the Twentieth Century Physics from a Historical Perspective

Chapter 1
Hermann Minkowski, Relativity and the Axiomatic Approach to Physics*

Leo Corry

Abstract This article surveys the general background to Minkowski's incursion into relativity, of which Einstein's work represented just one side. Special attention is paid to the idiosyncratic, rich, and complex interaction between mathematics and physics, that stood at the center of attention of the Gttingen mathematicians since the turn of the twentieth century. In particular the article explains Minkowski's formulation of special relativity in terms of space-time against the background of David Hilbert's program for the axiomatization of physics. In addition, the article sheds light on the changing attitudes of Einstein towards mathematics, in the wake of Minkowski's work, and his increasing willingness to attribute significance to mathematical formalism in developing physical theories.

Keywords Minkowksi · Hilbert · Axiomatization · Relativity · Gravitation

1.1 Introduction

In the history of both the special and the general theories of relativity two of the leading Göttingen mathematicians of the early twentieth century play a significant role: Hermann Minkowski (1864–1909) and David Hilbert (1862–1943). Although Minkowski and Hilbert accomplished their most important achievements in pure mathematical fields, their respective contributions to relativity should in no sense be seen as merely occasional excursions into the field of theoretical physics. Minkowski and Hilbert were motivated by much more than a desire to apply their exceptional mathematical abilities opportunistically, jumping onto the bandwagon of ongoing physical research by solving mathematical problems that physicists were unable to. On the contrary, Minkowski's and Hilbert's contributions to relativity are best understood as an organic part of their overall scientific careers.

L. Corry (✉)
Cohn Institute for History and Philosophy of Science, Tel-Aviv University
e-mail: corry@post.tau.ac.il

*This chapter is an adapted version of Chapter 4 of [5].

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_1,

Indeed, a detailed examination of their careers makes it evident that a keen interest in physics was hardly ever distant from either Hilbert's or Minkowski's main focus of activity in pure mathematics.[1] Minkowski's active interest in physics dates back at least to his Bonn years (1885–1894), during which he was in close contact with Heinrich Hertz (1857–1894). In 1888 he published an article on hydrodynamics in the proceedings of the Berlin Academy [29]. From his correspondence with Hilbert, we know that during his Zürich years (1896–1902) Minkowski kept alive his interest in mathematical physics, and in particular in thermodynamics. In 1902 he moved to Göttingen, following Hilbert's strong pressure on Felix Klein (1849–1925) to create a professorship for his friend. It is well known that during his last years there, Minkowski's efforts were intensively dedicated to electrodynamics. But this was not the only field of physics to which his attention was attracted. Minkowski was commissioned to write an article on capillarity [30] for the physics volume of the *Encyclopädie der mathematischen Wissenschaften*, edited by Arnold Sommerfeld (1868–1951). At several meetings of the Göttingen Mathematical Society he lectured on this, as well as on other physical issues such as Euler's equations of hydrodynamics and Nernst's work on thermodynamics, and the evolution of the theory of radiation through the works of Lorentz, Rayleigh, W. Wien, and Planck.[2] He also taught advanced seminars on physical topics and more basic courses on continuum mechanics, and gave exercises in mechanics and heat radiation.[3]

Like Minkowski, also Hilbert developed a strong interest in physics from very early on. Throughout his career he followed the latest developments closely and taught courses and seminars on almost every current physical topic. Hilbert elaborated the principles of his axiomatic method between 1894 and 1899 as part of his current interest in problems related to the foundations of geometry, but to a considerable extent, he also reflected throughout these years on the relevance of the method for improving the current state of physical theories. Influenced by his reading of Hertz's *Principles of Mechanics*, Hilbert believed that physicists often tended to solve disagreements between existing theories and newly found facts of experience by adding new hypotheses, often without thoroughly examining whether such hypotheses accorded with the logical structure of the existing theories they were meant to improve. In many cases, he thought, this had led to problematic situations in science which could be corrected with the help of an axiomatic analysis of the kind he had masterfully performed for geometry.[4] In a course taught in Göttingen in 1905 on the logical principles of mathematics, Hilbert gave a detailed overview of how such an axiomatic analysis would proceed in the case of several specific theories, including mechanics, thermodynamics, kinetic theory of gases,

[1] For details, see [5, pp. 11–25].

[2] As registered in the *Jahresbericht der Deutschen Mathematiker-Vereinigung* (*JDMV*). See Vol. 12 (1903), 445 & 447; Vol. 15 (1906), 407; Vol. 16 (1907), 78.

[3] See the announcement of his courses in JDMV Vol. 13 (1904), 492; Vol. 16 (1907), 171; Vol. 17 (1908), 116.

[4] See [5, pp. 83–110].

electrodynamics, probabilities, insurance mathematics and psychophysics.[5] In 1905 Hilbert and Minkowski, together with other Göttingen professors, organized an advanced seminar that studied recent progress in the theories of the electron.[6] In 1907, the two conducted a joint seminar on the equations of electrodynamics.[7] In the following sections I will argue that Minkowski's work can be seen to a large extent as a particular implementation of Hilbert's program for the axiomatization of physical theories, whereby the specific, structural role of a new principle recently adopted in various physical theories – the principle of relativity – was thoroughly investigated for the first time.

Albert Einstein (1879–1955) published his famous paper on the electrodynamics of moving bodies in 1905. Minkowski read this paper at some point, as did most of his colleagues at Göttingen. We know, for instance, that in October of 1907, Minkowski wrote to Einstein asking for a reprint, in order to study it in his joint seminar with Hilbert that semester.[8] Most likely, however, a much more direct and compelling source for his keen interest in the principle of relativity and its role in physics at large stemmed directly from his reading of the famous article on the dynamics of the electron, published by Henri Poincaré (1854–1912) in January of 1906.[9] For mathematicians at Göttingen it was routine to study attentively recent work published by Poincaré in all fields of research[10] and probably Minkowski and Hilbert were in a better position than anyone else to understand the breadth and the importance of these contributions, including his 1906 article. At the same time, Minkowski did not have a high appreciation of the mathematical abilities of Einstein (who studied in his courses at Zürich). He may also have been yet unaware of the profound impact of Einstein's work on leading theoretical physicists.[11]

---

[5] For details on this course, see [5, pp. 138–178].

[6] See [5, pp. 127–138].

[7] A detailed list of Hilbert courses on physics appears in [5], Appendix 2.

[8] Minkowski to Einstein, October 9, 1907 (*The Collected Papers of Albert Einstein* [*CPAE*] 2, Doc. 62).

[9] [42].

[10] Thus, for instance the *JDMV* mentions reports presented at the Göttingen Mathematical Society (GMG) (some of them by Minkowski himself) on Poincaré's recent works on probability, differential equations, capillarity, mathematical physics, topology, automorphic functions, boundary-value problems, function theory, and the uniformization theorem. Cf. *JDMV* 14 (1905), 586; 15 (1906), 154–155; 17 (1907), 5.

[11] The relative interest of Minkowski and his Göttingen colleagues in Poincaré's and Einstein's respective works as possible sources of information or inspiration on the topic has, of course, nothing to do with the question of priority between these two scientists concerning the "creation of the special theory of relativity". This more general, and perhaps abstract, question, that has attracted considerable attention from historians, is rather irrelevant for our account here. For a recent discussion of this topic, that emphatically attributes priority to Poincaré and at the same times provides a rather comprehensive list of references to the existing second literature see [13] (Also available at http://albinoni.brera.unimi.it/Atti-Como-98/Giannetto.pdf). For a more recent account of Poincaré's work in relativity and its background, see [20].

Beginning in 1907, at any rate, Minkowski erected the new theory of relativity on what was to become its standard mathematical formulation, and he also devised the language in which it was further investigated. In particular, Einstein's adoption of Minkowski's formulation – after an initial unsympathetic attitude towards it – proved essential to his own attempts to generalize the theory so that it would cover gravitation and arbitrarily accelerated systems of reference. Minkowski's ideas concerning the postulate of relativity have been preserved in the manuscript and published versions of three public talks, as well as through an article posthumously published by Max Born (1882–1970), based on Minkowski's papers and on conversations between the two. The first public presentation of these ideas took place in November 5, 1907, in a talk delivered to the GMG under the name of "The Principle of Relativity,"[12] barely 1 month after requesting Einstein's paper.

Attempts to deal with the electrodynamics of moving bodies since the late nineteenth century had traditionally comprised two different perspectives: the microscopic theories of the electron and the macroscopic, or phenomenological, theories of optical and electromagnetic phenomena in moving media.[13] Whereas Einstein's 1905 relativistic kinematics concerned only Lorentz's microscopic electron theory, it was Minkowski who first addressed the problem of formulating a phenomenological relativistic electrodynamics of moving media. Thus his three public lectures on the postulate of relativity deal mainly with the macroscopic perspective, while the application of his point of view to addressing the microscopic perspective appeared in the posthumous article published by Born.

In the historiography of relativity theory, Minkowski's contributions to this domain were often judged, as were those of most of his contemporaries, against his perceived ability to understand the impact of Einstein's innovations.[14] This led to a remarkable oversight of his well-known collaboration with Hilbert as an important factor to be considered in describing and explaining his incursion into relativity theory.[15] More recent studies have adopted a broader perspective and have helped

[12] Published as [34]. For details on the printed and manuscript versions of Minkowski's work see [12, pp. 119–121]. The original typescript of this lecture was edited for publication by Sommerfeld. After comparing the published version with the original typescript, Lewis Pyenson [44, pp. 82] has remarked that Sommerfeld introduced a few changes, among them a significant one concerning the role of Einstein: "Sommerfeld was unable to resist rewriting Minkowski's judgment of Einstein's formulation of the principle of relativity. He introduced a clause inappropriately praising Einstein for having used the Michelson experiment to demonstrate that the concept of absolute space did not express a property of phenomena. Sommerfeld also suppressed Minkowski's conclusion, where Einstein was portrayed as the clarifier, but by no means as the principal expositor, of the principle of relativity." The added clause is quoted in [12, pp. 93].

[13] On the development of these two perspectives before Einstein and Minkowski, see *CPAE* 2, 503–504.

[14] Cf., e.g., [46, pp. 144]: "Hermann Minkowski, the mathematician who used Einstein's special theory of relativity to elaborate during the years 1907–1909 a theory of absolute, four-dimensional space-time ... understood little of Einstein's work and his main objective lay in imposing mathematical order on recalcitrant physical laws."

[15] For example, no such connection was considered in previous, oft-cited accounts of Minkowski's work: [12; 44; 27, pp. 238–244].

understand the immediate framework of scientific interests of Minkowski and to explain how these works fit therein, not just as a side issue to the main story of Einstein's development of the theory of relativity.[16]

In the present chapter I explain how the newly introduced relativistic ideas were combined by Minkowski with ideas embodied in Hilbert's program of axiomatization. This interpretation helps understanding the motivations and actual scope of his work and at the same time it also stresses the kind of questions that Minkowski was *not* pursuing in his work. In particular, the point of view adopted here suggests a reinterpretation of the role of Minkowski's work in the debates of the first decade of the century – much discussed in the secondary literature – concerning the ultimate nature of natural phenomena. In the earlier historiography, Minkowski's work was often presented as an attempt to elaborate and support the so-called "electromagnetic worldview" as a foundational position in physics opposed to mechanistic reductionism.[17] This debate, in which various physicists participated with varying degrees of intensity at the turn of the twentieth century, appears as irrelevant to my presentation of Minkowski's work.

1.2 The Principle of Relativity

Minkowski's first talk on electrodynamics at the meeting of the GMG in November 1907 was basically a direct continuation of his recent joint seminar with Hilbert, where they had also studied Einstein's 1905 paper. We have limited information about this seminar,[18] but we do know that in one of its meetings Hilbert discussed the electrodynamics of moving bodies. Hilbert described geometrical space as being filled with three different kinds of continua: ether, electricity and matter. The properties of these continua, he said, should be characterized by suitable differential equations. Thus the ether, a medium at rest, is characterized in terms of the magnetic and electric field intensities, $\boldsymbol{M}$ and $\boldsymbol{e}$ respectively. Electricity, a medium in motion, is characterized in terms of the current density vector and the scalar charge density, $\boldsymbol{s}$ and ρ respectively.[19] A main task of electrodynamics, Hilbert stated, is the determination of the latter two magnitudes in the presence of external forces. Hilbert seems to have expressed doubts concerning the adequacy of Lorentz's equations to describe the electrodynamics of moving bodies. At any rate, the equations discussed in the seminar were those on which Minkowski based his talk, albeit using his innovative formulation in terms of four-vectors.

[16] [55–58].

[17] See [7, Chap. 9; 19, pp. 231–242].

[18] Notes of the seminar were taken by Hermann Mierendorff, and they are preserved at the David Hilbert *Nachlass* in Göttingen (*DHN* 570/5). Cf. [44, pp. 83], for additional details.

[19] For the sake of uniformity throughout the forthcoming sections I have slightly modified the original notation and symbols. These changes are minor and should not produce interpretive problems, though. On this important point see [58].

Minkowski opened his talk by declaring that recent developments in the electromagnetic theory of light had given rise to a completely new conception of space and time, namely, as a four-dimensional, non-Euclidean manifold. Whereas physicists were still struggling with the new concepts of the theory painfully trying to find their way through the "primeval forest of obscurities," mathematicians have long possessed the concepts with which to clarify this new picture. At the center of these developments lies the principle of relativity. The impact of these developments had created a state of great conceptual confusion in many physical disciplines. The aim of Minkowski's new investigations was to clarify, to understand and to simplify the conceptual edifice of electrodynamics and mechanics, while sorting out the fundamental statements – including the principle of relativity – that lie at the basis of those disciplines. The implications derived from these first principles had to be confronted by experiment in order to validate or refute the relevant theories. Minkowski introduced here many of the mathematical concepts and terms that have come to be associated with his name and that became standard in any discussion of relativity, but he did not treat them systematically at this stage.

Minkowski was not speaking specifically about Einstein and about his 1905 paper, but rather about a broader trend that included the work of Lorentz, FitzGerald, Poincaré, and Planck. A proper elaboration of their ideas, he said, could become one of the most significant triumphs in applying mathematics to understanding the world, provided – he immediately qualified his assertion – "they actually describe the observable phenomena."[20] This latter, brief remark characterizes very aptly the nature of Minkowski's incursion into the study of the electrodynamics of moving bodies: along the lines of Hilbert's analysis of the axioms of other physical disciplines, he would attempt to understand and simplify the conceptual structures of electrodynamics and mechanics – presently in a state of great confusion, in view of the latest discoveries of physics. He would sort out the fundamental statements that lie at the basis of those structures, statements that must be confronted by experiment in order to validate or refute the relevant theories. The fundamental role played by the principle of relativity would thus be clarified.

Minkowski's main technical innovation consisted in introducing the magnitudes of four and of six components (he called the latter "*Traktoren*"), together with a matrix calculus, as the mathematical tools needed to bring to light all the symmetries underlying relativistic electrodynamics.[21] Minkowski claimed that the four-vector formulation reveals the full extent of the invariance properties characteristic of Lorentz's equations for the electron. It took a mathematician of the caliber of Minkowski to recognize the importance of Poincaré's group-theoretical interpretation of the Lorentz transformations, but he also pointed out that earlier authors, like Poincaré, had not previously emphasized that the equations satisfy this kind of purely formal property, which his newly introduced formalism made

[20] [34, pp. 927]: "falls sie tatsächlich die Erscheinungen richtig wiedergeben, . . ."

[21] For the place of Minkowski's contribution in the development of the theory of tensors, see [48, pp. 168–184]. The term "four-vector" was introduced in [53].

quite evident.[22] In this earliest presentation Minkowski did not actually write down the Maxwell equations in manifestly Lorentz-covariant form. Still, he showed sketchily that if the quantities that enter the equations are written in terms of four-vectors, their invariance under any transformation that leaves invariant the expression ${x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2$ (where $x_4 = it$) follows as a simple mathematical result. Thus formulated, the Lorentz transformations represent rotations in this four-dimensional space.

Minkowski stressed that his theory does not assume any particular worldview as part of a foundational position in physics: it treats first electrodynamics and only later mechanics, and its starting point is the assumption that the correct equations of physics are still not entirely known to us. Perhaps 1 day a reduction of the theory of matter to the theory of electricity might be possible, Minkowski said, but at this stage only one thing was clear: experimental results, especially the Michelson experiment, had shown that the concept of absolute rest corresponds to no property of the observed phenomena. He proposed to clarify this situation by assuming that the equations of electrodynamics remain invariant under the Lorentz group even *after* matter had been added to the pure field. Precisely here the principle of relativity enters the picture of physics, for Minkowski declared that this principle – i.e., invariance under Lorentz transformations – is a truly new kind of physical law: Rather than having been deduced from observations, *it is a demand we impose on yet to be found equations describing observable phenomena*.[23]

Minkowski used the four-vector formulation to show how the Galilean mechanics arises as a limiting case when $c = \infty$. Similarly, he derived the electrodynamic equations of a moving medium, making evident and stressing their invariance under the Lorentz group. He thus concluded that if the principle of relativity is to be valid also for matter in motion, then the basic laws of classical mechanics could only be approximately true. The impossibility of detecting the motion of the earth relative to the ether (following the Michelson experiment) thus implies the validity of the relativity principle.[24] As a further argument to support this rejection the classical principle of inertia Minkowski also quoted an elaborate technical argument taken from Planck's recent contribution to a relativistic thermodynamics.[25]

Minkowski concluded his lecture with a brief discussion on gravitation. Naturally, if the principle of relativity was to be truly universal it should account also for phenomena of this kind. Minkowski mentioned a similar discussion that had appeared in Poincaré's relativity article, and endorsed Poincaré's conclusion there that gravitation must propagate with the velocity of light. The purely mathematical task thus remained open, to formulate a law that complies with the relativity principle, and at the same time has the Newtonian law as its limiting case. Poincaré had indeed introduced one such law, but Minkowski regarded this law as only one among many

[22] [34, p. 929].

[23] [34, p. 931].

[24] [34, pp. 932–933].

[25] [34, pp. 935–937]. He referred to [39]. For an account of Planck's paper, see [27, pp. 360–362].

possibilities, noting that Poincaré's results had hitherto been far from conclusive. At this early stage of development of relativistic thinking in physics, the general perception was that the incorporation of Newtonian gravitation would pose only minor problems.[26] This also seems to have been Minkowski's opinion, and he left the more elaborate treatment of this point for a later occasion. Of course, he could not have imagined at this point how elusive and difficult this task would turn out to be.[27]

1.3 The Basic Equations of Electromagnetic Processes in Moving Bodies

Minkowski's second talk, "The Basic Equations of Electromagnetic Processes in Moving Bodies", was his only published text on this topic to appear before his death in 1909.[28] The talk was delivered at the meeting of the Göttingen Scientific Society (GWG) on December 21, 1907, only 2 weeks after Klein had lectured at the GMG on the possible applications of the quaternion calculus to the theory of the electron and its relation to the principle of relativity. Following Klein's lecture, Minkowski showed how the equations of electrodynamics can be simplified if the electric and magnetic magnitudes are jointly represented by means of bi-quaternions, namely, quaternions with complex components, and how this is related to the study of the significance of the principle of relativity.[29]

Minkowski's talk contained his most detailed mathematical treatment of the differential equations of electrodynamics. It also presented an illuminating conceptual analysis, very similar in spirit to Hilbert's axiomatic treatment of physical theories, of the main ideas involved in the current developments of the theories of the electron and of the role played by the principle of relativity in those theories. It is therefore not surprising that Hilbert considered this talk to be his friend's most significant contribution to electrodynamics. In his obituary of Minkowski, Hilbert stressed the importance and innovative character of the axiomatic analysis presented in that article, especially for Minkowski's derivation of the equations for moving matter starting from the so-called "World-postulate" and three additional axioms. The correct form of these equations had been theretofore a highly controversial issue among physicists, but this situation had totally changed – so Hilbert believed – thanks to Minkowski's work.[30]

---

[26] Cf. [36, pp. 20–21].

[27] Cf. [58] for additional details.

[28] [32].

[29] See the announcement in *JDMV* 17 (1908), 5–6.

[30] [17, pp. 93–94].

1.3.1 Three Meanings of "Relativity"

Minkowski based his conceptual analysis on a clear distinction between three possible different meanings that may be associated with the principle of relativity. First, there is the plain mathematical fact that the Maxwell equations, as formulated in Lorentz's theory of electrodynamics, are invariant under the Lorentz transformations. Minkowski called this fact the "*theorem* of relativity." Second, it seemed natural to expect, that the domain of validity of the theorem – a mathematically evident theorem, in his opinion – might be extended to cover *all* laws governing ponderable bodies, including laws that are still unknown. This is the "*postulate* of relativity," which expresses a confidence (*Zuversicht*) rather than an objective assessment concerning about the actual state of affairs. One can embrace this confidence, claim, Minkowski stressed, *without thereby committing oneself to any particular view of the ultimate relationship between electricity and matter*.[31] He compared this postulate to the principle of conservation of energy, which we assume even for forms of energy that are not yet known. Lastly, if we can assert that the expected Lorentz covariance actually holds as a relation between directly observable magnitudes relating to a moving body, then this particular relation is called the "*principle* of relativity."

From Minkowski's analysis of these three distinct interpretations of the notion of relativity we can also learn about his views on the specific contributions of the various physicists to the topics discussed. Thus, Lorentz had discovered the theorem and had also set up the postulate of relativity in the form of the contraction hypothesis. Einstein's contribution was, according to Minkowski, that of having very clearly claimed that the postulate (of relativity) is not an artificial hypothesis, but rather, that the observable phenomena force this idea upon us as part of a new conception of time. Minkowski did not mention Poincaré by name, but given the latter's conception of the general validity of the theorem, he would presumably have classified Poincaré's contribution as having also formulated the "relativity postulate." In fact, it was Poincaré who first suggested that the domain of validity of Lorentz invariance should be extended to all domains of physics. In 1904, for instance, he formulated the principle as an empirical truth, still to be confirmed or refuted by experiment, according to which the laws of physics should be the same for any two observers moving with rectilinear, uniform motion relative to each other.[32]

These attributions of his predecessors achievements served to support Minkowski's claim that his interpretation of the principle of relativity for the electrodynamics of moving bodies was a novel approach. His presentation aimed to deduce an exact formulation of the equations of moving bodies from the principle

[31] [32, pp. 353] (emphasis added).

[32] [41, p. 495; 42, p. 176]. And again in [43, p. 221]: "It is impossible to escape the impression that the Principle of Relativity is a general law of nature. . . . It is well [sic] in any case to see what are the consequences to which this point of view would lead, and then submit these consequences to the test of experiment."

of relativity, thus making clear that none of the existing formulations was fully compatible with the principle. Minkowski believed that his axiomatic interpretation of the principle of relativity was the best approach for unequivocally obtaining the correct equations. Furthermore, the invariance of these equations under the Lorentz group would follow from simple symmetry considerations.[33]

In a separate section Minkowski discussed the changes in our concepts of time implied by the introduction of the Lorentz transformations into kinematics, and in particular the impossibility of speaking about the simultaneity of two events. This section may have drawn some inspiration from a well-known article of 1906 by Kaufmann.[34] In a lengthy review of all recent experiments for testing the theories of the electron, Kaufmann established that his own results were incompatible with the "Lorentz-Einstein approach", an approach he also rejected because it did not comply with the electromagnetic world-view, which Kaufmann staunchly supported. This article attracted considerable attention, including a detailed critique by Planck, which offered open, if cautious, support for a continued study of relativity and its consequences for physics.[35] Kaufmann attributed to Einstein a new derivation of the electromagnetic equations for moving bodies in which the principle of relativity was placed at the foundation of all physical theories. In addition, he attributed to Einstein the introduction of a new conception of time that dispensed with the concept of simultaneity for two separate points in space. In his rebuttal, Planck asserted that Lorentz had introduced the principle of relativity and Einstein had formulated a much more general version of it. These two articles, which Minkowski undoubtedly read, were part of a longer series of early historical accounts that started appearing alongside the early development of the theory itself. These created different conceptions of the specific contributions of the various scientists involved.[36]

It is also noteworthy that this section appears at the end of Minkowski's discussion of the equations in empty ether. Clearly, he saw the relativity of simultaneity as a consequence of the Lorentz theorem for the equations for the ether, and thus as a fact independent of the ultimate nature of matter. The relativity of simultaneity, Minkowski moreover thought, should not pose particular difficulties to mathematicians. Familiar as the latter were with higher-dimensional manifolds and non-Euclidean geometries, they should easily adapt their concept of time to the new one. On the other hand, Minkowski noted that the task of making physical sense of the Lorentz transformations should be left to physicists, and in fact he saw the introduction of Einstein's 1905 relativity article as attempting to fulfill this task.[37]

---

[33] Minkowski formulated this statement in terms of four-vectors of four and six components (which he called "space-time vectors of type I and II", respectively). Vectors of type II correspond to modern second-rank, antisymmetric tensors.

[34] [21].

[35] [39]. Cf. [14, pp. 28–31].

[36] Cf. [55].

[37] [32, p. 362].

1.3.2 Axioms of Electrodynamics

Minkowski devoted a long section to analyzing in detail the Maxwell-Lorentz equations together with the underlying axioms of the theory. This section is of special interest for our purposes here, since it clearly brings to the fore the close connections between Minkowski's and Hilbert's ideas in this domain. The starting point was Lorentz's version of Maxwell's equations for the case of matter at rest in the ether, which Minkowski formulated as follows:

$$\mathrm{curl}\boldsymbol{m} - \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{t}} = \boldsymbol{s} \tag{1.1}$$

$$\mathrm{div}\boldsymbol{e} = \rho \tag{1.2}$$

$$\mathrm{curl}\boldsymbol{E} + \frac{\partial \boldsymbol{M}}{\partial \boldsymbol{t}} = 0 \tag{1.3}$$

$$\mathrm{div}\boldsymbol{M} = 0 \tag{1.4}$$

$\boldsymbol{M}$ and $\boldsymbol{e}$ are called the magnetic and electric intensities (*Erregung*) respectively, $\boldsymbol{E}$ and $\boldsymbol{m}$ are called the electric and magnetic forces, ρ is the electric density, $\boldsymbol{s}$ is the electric current vector (*elektrischer Strom*). Further, Minkowski limited his discussion to the case of isotropic bodies by adding three conditions that characterize matter in this case:

$$\boldsymbol{e} = \varepsilon \boldsymbol{E}, \quad \boldsymbol{M} = \mu \boldsymbol{m}, \quad \boldsymbol{s} = \sigma \boldsymbol{E}, \tag{1.5}$$

where ε is the dielectric constant, μ is the magnetic permeability, and σ is the conductivity of matter.

Minkowski sought to derive now the equations for matter in motion, and in doing so he followed and approach that strongly reminds the procedures suggested by Hilbert in his axiomatization lectures, although the details of the implementation are much more elaborated in this case than they were in any of Hilbert's presentations so far. To the equations for matter at rest Minkowski added three axioms meant to characterize the specific physical situation in mathematical terms. Thus, the three axioms are:

1. Whenever the velocity $\boldsymbol{v}$ of a particle of matter equals 0 at x, y, z, it in some reference system, then Eqs. (1.1–1.5) also represent, in that system, the relations among all the magnitudes: ρ, the vectors $\boldsymbol{s}$, $\boldsymbol{m}$, $\boldsymbol{e}$, $\boldsymbol{M}$, $\boldsymbol{E}$, and their derivatives with respect to x, y, z, it.
2. Matter always moves with a velocity which is less than the velocity of light in empty space (i.e., $|\boldsymbol{v}| = v < 1$).
3. If a Lorentz transformationacting on the variables x, y, z, it, transforms both $\boldsymbol{m}$, $-i\boldsymbol{e}$ and $\boldsymbol{M}$, $-i\boldsymbol{E}$ as space-time vectors of type II, and $\boldsymbol{s}$, $i\rho$ as a space-time vector of type I, then *it* transforms the original equations exactly into the same equations written for the transformed magnitudes.[38]

[38] [32, p. 369]. For the sake of simplicity, my formulation here is slightly different but essentially equivalent to the original one.

Minkowski called this last axiom, which expresses in a precise way the requirement of Lorentz covariance for the basic equations of the electrodynamics of moving matter, the principle of relativity. It is relevant to see in some detail how Minkowski applies the axioms to derive the equations.

Since $v < 1$ (axiom 2), Minkowski could apply a result obtained in the first part, according to which the vector **v** can be put in a one-to-one relation with the quadruple

$$w_1 = \frac{v_x}{\sqrt{1-v^2}}, \quad w_2 = \frac{v_y}{\sqrt{1-v^2}}, \quad w_3 = \frac{v_z}{\sqrt{1-v^2}}, \quad w_4 = \frac{i}{\sqrt{1-v^2}}$$

which satisfies the following relation:

$$w_1^2 + w_2^2 + w_3^2 + w_4^2 = -1.$$

Again from the results of the first part, it follows that this quadruple transforms as a space-time vector of type I. Minkowski called it the "velocity space-time-vector." Now, if $\boldsymbol{v} = 0$, by axiom 1, Eqs. (1.1–1.5) are also valid for this case. If $\boldsymbol{v} \neq 0$, since $|\boldsymbol{v}| < 1$, again the results of earlier sections allow the introduction of a transformation for which

$$w_1{}' = 0, \ w_2{}' = 0, \ w_3{}' = 0, \ w_4{}' = i.$$

In this case, we also obtain a transformed velocity $\boldsymbol{v}' = 0$. According to axiom 3, whatever the basic equations may be that hold for this case must remain invariant when written for the transformed variables x', y', z', t' and the transformed magnitudes $\boldsymbol{M}', \boldsymbol{e}', \boldsymbol{E}', \boldsymbol{m}', \rho', \boldsymbol{s}'$, and the derivatives of the latter with respect to x', y', z', t'. But, since $\boldsymbol{v}' = 0$, the transformed equations are (by axiom 1) just (1.1'–1.4'), obtained from (1.1–1.4) by tagging all variables. The same is true for Eq. (1.5) (although there is no need to apply axiom 3), but with ε, μ, and σ remaining unchanged. Finally, one applies the inverse of the original Lorentz transformation and, by axiom 3, it follows that the form of the basic equations for the original variables is in fact precisely that of (1.1–1.4). Minkowski thus concluded that the basic equations of electrodynamics for moving bodies are the same as the equations for stationary bodies, and the effects of the velocity of matter are manifest only through those conditions in which its characteristic constants ε, μ, and σ appear. Also, Minkowski concluded, the transformed Eq. (1.5′) can be transformed back into the original Eq. (1.5).

The arguments advanced in this section are quite different from the elaborate mathematical and physical arguments displayed throughout much of Minkowski's talk, and, at first sight, they may appear as somewhat out of place here. However, when seen in the light of the kind of axiomatic conceptual clarification promoted by Hilbert in his lectures on physics, they would seem to find a more natural place. In fact, still under the same perspective, Minkowski proceeded to check if, and to what extent, alternative, existing versions of the equations also might satisfy the principle of relativity, as formulated in his axioms. The implicit assumption was that only equations consistent with his version of the principle of relativity could be accepted as correct. Minkowski thus found, for instance, that the macroscopic equations for moving media formulated in 1904 by Lorentz were incompatible with his principle in certain cases.[39] Likewise, the equations formulated in 1902 by Emil Cohn (1854–1944) agreed with Minkowski's own, up to terms of first order in the velocity.[40] This was a point of major significance for Minkowski. In the introduction to his article he had pointed out that, perhaps surprisingly, Lorentz's own equations for moving bodies did not correspond to the principle of relativity, and thus a major task of his article would be the formulation of the appropriate, invariant equations. In doing so, he was drawing a then unprecedented, and certainly important, distinction between Lorentz's theory of the electron and the consequences of relativity.[41] As my account here shows, this important task was reached by relying precisely on the axiomatic analysis of the theory and the principle of relativity.

1.3.3 Relativity and Mechanics

Three additional sections of this paper discuss the properties of electromagnetic processes in the presence of matter, while an appendix discusses the relations between mechanics and the postulate (not the principle!) of relativity. It is here that the similarity between Minkowski's and Hilbert's treatments of physical theories becomes most clearly manifest. Hilbert had spoken many times in the recent past about the frequent situation in the history of physics wherein new hypotheses were added to existing theories only on the basis of their intrinsic plausibility and without thoroughly checking if the former contradict the latter or any of their direct consequences. One of Hilbert's expressed aims in applying the axiomatic method to physical theories was to avoid such potential pitfalls. And indeed, it was precisely in order to avoid the danger of such a possible contradiction in the framework of the recent, exciting developments in physics that Minkowski undertook this painstaking conceptual analysis of the ideas involved. In this final section, he explored in detail the consequences of adding the postulate of relativity to the existing edifice of mechanics, as well as its compatibility with the already established principles of the discipline. The extent to which this could be successfully realized would provide a standard for assessing the status of Lorentz covariance as a truly universal postulate for all physical science.

Using the formalism developed in the earlier sections Minkowski showed that in order for the equations of motion of classical mechanicsto remain invariant under the Lorentz groupit is necessary to assume that $c = \infty$. It would be embarrass-

[39] [32, p. 372]. The article is [25].

[40] Minkowski cited here [4]. For Cohn's electrodynamics see [6, pp. 271–276].

[41] Cf. [55, footnote 15]

ing or perplexing (*verwirrend*), he said, if the laws of transformation of the basic expression

$$-x^2 - y^2 - z^2 + c^2t^2$$

into itself were to necessitate a certain finite value of c in a certain domain of physics and a different, infinite one, in a second domain. Accordingly, the postulate of relativity (i.e., our confidence in the universal validity of the theorem) compels us to see Newtonian mechanics only as a tentative approximation initially suggested by experience, which must then be corrected to make it invariant for a finite value of c. Minkowski not only thought that reformulating mechanics in this direction was possible (he asserted) in terms very similar to those found in Hilbert's lecture notes, that such a reformulation seemed to add substantially to the perfection of the axiomatic structure of mechanics.[42]

Naturally, the discussion in this section was couched in the language of space-time coordinates x, y, z, t. But Minkowski referred throughout to the properties of matter at a certain point of *space* at a given *time*, clearly separating the three elements, and focusing on the path traversed by a particle of matter throughout time. The space-time line is the collection of all the space-time points x, y, z, t associated with that particle, and the task of studying the motion of matter is defined as follows: "For every space-time point to determine the direction of the space-time line traversed by it." Likewise, the collection of all space-time lines associated with the material points of an extended body is called its space-time thread (*Raum-Zeitfaden*). One can also define the "proper time" of a given matter particle in these terms, generalizing Lorentz's concept of local time, and one can associate a positive magnitude (called *mass*) to any well-delimited portion of (three-dimensional!) space at a given time. These last two concepts lead to the definition of a rest-mass density, which Minkowski used to formulate the principle of conservation of mass. Thus, Minkowski relied here on the four-dimensional language as an effective, *formal* mathematical tool providing a very concise and symmetric means of expression, rather than as a new, *intuitive* geometrical understanding of space-time. The innovative conception usually attributed to Minkowski in this regard would only appear fully articulated in his talk of 1908 in Köln (discussed below).

Still using the same language, Minkowski analyzed the compatibility of the world-postulate with two accepted, basic principles of mechanics: Hamilton's principle and the principle of conservation of energy. He stressed with particular emphasis the full symmetry with respect to all four variables x, y, z, t, for the equations obtained. Integrating the terms of the equations of motion that had been derived by means of the Hamilton principle, he obtained four new differential equations

$$m\frac{d}{d\tau}\frac{dx}{d\tau} = R_x,$$
$$m\frac{d}{d\tau}\frac{dy}{d\tau} = R_y,$$

[42] [32, p. 393].

$$m\frac{d}{d\tau}\frac{dz}{d\tau} = R_z,$$
$$m\frac{d}{d\tau}\frac{dt}{d\tau} = R_t.$$

Here m is the constant mass of a thread, τ is the proper time, and R is a vector of type I: the *moving force* of the material points involved. The full symmetry obtained here by the adoption of the postulate of relativity struck Minkowski as highly significant, especially concerning the status of the fourth equation. Echoing once again the spirit and the rhetoric of Hilbert's lectures on axiomatization he concluded that this derivation, which he deemed surprising, entirely justifies the assertion that if the postulate of relativity is placed at the foundations of the building of mechanics, the equations of motion can be fully derived from the principle of conservation of energy alone.[43]

1.3.4 Relativity and Gravitation

Minkowski's brief treatment of gravitation follows a similar rationale: it should be proved that the World-postulate does not contradict the relevant, observable phenomena, and where necessary, the existing theory has to be suitably reformulated. Obviously, the truly universal validity of the postulate could only be asserted if it covered this domain as well, which was traditionally considered to be particularly problematic. Thus, in the closing passages, Minkowski sketched his proposal for a Lorentz-covariant theory of gravitation, much more elaborate than the one presented in his previous talk. A brief description of this section is relevant here since the general principles of the approach followed by Minkowski in developing his gravitational considerations are closely related with those of Hilbert later on. It is also noteworthy that in this section Minkowski elaborated his four-dimensional formulation even further, introducing ideas quite close to the notion of a light cone and the kind of reasoning associated with it. In this regard the overall approach of this section on gravitation can be described as much more geometric, in the basic, visual-intuitive sense of the term (albeit in four dimensions rather than the usual three), than all previous ones dealing with electrodynamics and even with mechanics.

In order to adapt Newton's theory of gravitation to the demand of Lorentz covariance Minkowski described in four-dimensional geometrical terms the force vector acting on a mass particle m at a certain point B. This vector has to be orthogonal to the world-line of the particle at B, since four-force vectors are orthogonal to four-velocity vectors. To remain close to Newton's theory, Minkowski also assumed that the magnitude of this vector is inversely proportional to the square of the distance (in ordinary space) between any two mass particles. Finally, he also assumed that the actual direction of the orthogonal vector to the world-line of m is in fact determined by the line connecting the two attracting particles. These requirements must all be

[43] [32, p. 401].

satisfied by any adaptation of Newton's laws to Lorentz covariance, but of course, Minkowski still had to be more specific in his choice of such a law. He did so in the following way: Take a fixed space-time point $B^*(x^*, y^*, z^*, t^*)$, and consider all the points $B(x, y, z, t)$ satisfying the equation

$$(x - x^*)^2 + (y - y^*)^2 + (z - z^*)^2 = (t - t^*)^2, \quad (t - t^* \geq 0).$$

This is called the "light-structure" of B^*, and B^* is a light-point in the set of all the points located towards the concave side of the three-surface defined by the light-structure. Using the language introduced later by Minkowski himself, one can say that B^* can communicate by light signals with all points of which it is a light-point. If in the above relation, B^* is taken as variable and B as fixed, then Minkowski claimed that for an arbitrarily given space-time line there exists only one point B^* which is a light-point of B. This latter conclusion is valid only if the space-time line is (using the terminology introduced later) time-like, which is implicit in Minkowski's definition of space-time lines as world-lines of matter.[44] Given two matter points F, F^* with masses m, m^*, respectively, assume F is at space-time point B, and let BC be the infinitesimal element of the space-time line through F. This space-time line is nothing but the (modern language) word-lines of the particles at those events, with masses m, m^*. Minkowski claimed that the moving force of the mass point F at B should (*möge*) be given by a space-time vector of type I, which is normal to BC, and which equals the sum of the vector described by the formula

$$mm^* \left(\frac{OA'}{B^*D^*} \right)^3 BD^*, \tag{1.6}$$

and a second, suitable vector, parallel to B^*C^*. Figure 1.1 may help clarifying Minkowski's train of thought. The additional space-time points that appear in the

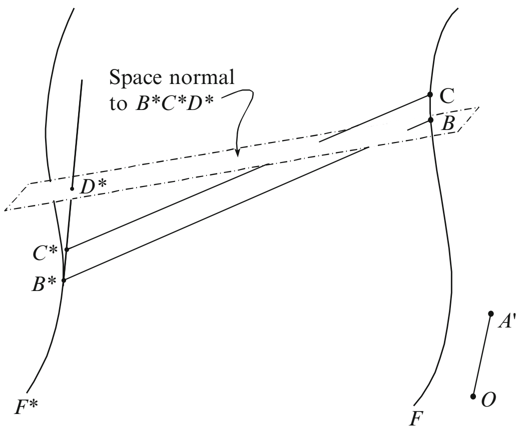


Fig. 1.1 A schematic representation of Minkowski's relativistic treatment of gravitation, for two matter points F, F^*

[44] [32, p. 393].

diagram are defined by Minkowski (without himself using any figure) as follows: B^* is the light-point of B along the space-time line of F^*; O is the origin of the coordinate system and OA' is a segment parallel to B^*C^* (C^* being the light-point along the world-line of F^*, of space-time point C) whose endpoint A' lies on the four-dimensional hyperbolic surface

$$-x^2 - y^2 - z^2 + t^2 = 1.$$

Finally, D* is the intersection point of the line through B^*C^* and the normal to OA' passing through B.

Minkowski added the assumption that the material point F^* moves uniformly, i.e., that F^* describes a straight line. Thus, at the outset he has presumably assumed that F^* moves arbitrarily. In this more general case, BC and B^*C^* represent the tangent vectors to the curves F and F^*, and they can be physically interpreted as the four-velocities of the masses with world-lines F and F^*, respectively. Now, Minkowski's gravitational force must be orthogonal to the four-velocity of F at B, and therefore orthogonal to BC. B^*C^*, on the other hand, helps to determine the distance between F and F^* in the rest-frame of the attracting body F^*, a magnitude necessary to make the gravitational law inversely proportional to it. In effect the velocity of F^* at B^* is parallel to B^*C^*, and by extending the latter into B^*D^*, Minkowski is determining the plane on which the desired distance should be measured, i.e., a plane which is normal to B^*D^* and passes through B. The space distance (not space-time) between the two points is thus given by BD^*.

Now the quantity BD^* also appears in Eq. (1.6) and in fact it gives the direction of the vector represented by the latter. But, as said above, the gravitational force should be orthogonal to BC, which is not necessarily the case for BD^*. Minkowski corrected this situation by adding to the first vector a second "suitable" one, parallel to B^*C^*. Thus the "suitable" vector that Minkowski was referring to here is one that, when added to Eq. (1.6) yields a third vector which is orthogonal to BC.

The product of the masses m and m^* appears in Eq. (1.6) and to that extent it directly corresponds to the Newtonian law. But does this equation really embody an inverse square law in the present situation? It seems that Minkowski's additional assumption, i.e., that F^* moves uniformly, could serve to answer this question (although Minkowski does not explicitly elaborate on this point). In fact, after this assumption is added, the new situation can be represented as in Fig. 1.2. If one sets the coordinates of B^* to be $(0, 0, 0, \tau^*)$, then the origin O lies on F^*. Moreover, the following values of the magnitudes involved in the equation can be deduced directly from their definitions:

$$OA' = 1; B^*D^* = t - \tau^*; (BD^*)^2 = x^2 + y^2 + z^2.$$

But B^* is a light point of B, and therefore

$$(B^*D^*)^2 = (t - \tau^*)^2 = x^2 + y^2 + z^2.$$

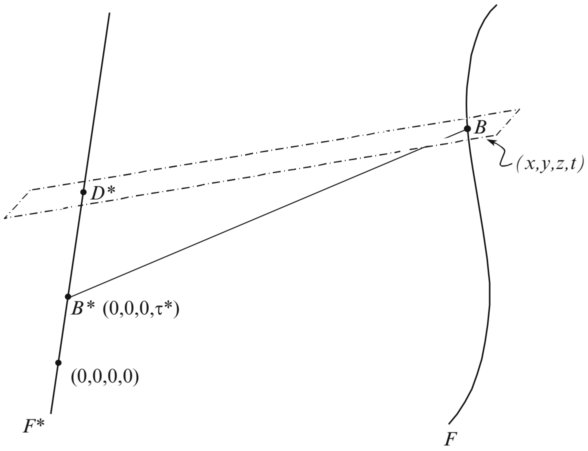


Fig. 1.2 A schematic representation of Minkowski's relativistic treatment of gravitation, with the body F^* moving uniformly

Equation (6) is thus reduced to the following:

$$mm^* \left(\frac{OA'}{B^* D^*} \right)^3 BD^* = - \frac{mm^*}{(x^2 + y^2 + z^2)},$$

which is the desired inverse square law of gravitation. Moreover, the assumption that F^* moves uniformly also prepares the way for Minkowski's discussion of the solar system at the end of his article (see below), by letting F^* represent the inertial motion of the sun and F the non-inertial motion of an orbiting planet.

Although many details of Minkowski's argument (such as those presented here) do not appear in the printed version of his article, all the discussion was fully conducted in the framework of space-time geometry, using only four-vectors defined on world-points and world-lines. Minkowski could thus conclude, without further comment, that the above determination of the value of the moving force is covariant with respect to the Lorentz group.

Minkowski went on to determine how the space-time thread of F behaves when the point F^* undergoes a uniform translatory motion. He asserted that starting from Eq. (1.6) as the value of the attracting force, the following four equations could be obtained:

$$\frac{d^2x}{d\tau^2} = - \frac{m^* x}{(t - \tau^*)^3}, \quad \frac{d^2y}{d\tau^2} = - \frac{m^* y}{(t - \tau^*)^3}, \quad \frac{d^2z}{d\tau^2} = - \frac{m^* z}{(t - \tau^*)^3}, \tag{1.7}$$

and

$$\frac{d^2t}{d\tau^2} = - \frac{m^* x}{(t - \tau^*)^2} \frac{d(t - \tau^*)}{dt}. \tag{1.8}$$

Since the relation $x^2 + y^2 + z^2 = (t - \tau^*)^2$ holds true, Eq. (1.7) is a set of equations similar to the motion equations of a material point under the Newtonian attraction of a fixed center, as Minkowski stated, substituting instead of the time t the proper time τ of the particle. On the other hand, Eq. (1.8) establishes the dependence between the proper time of the particle and the time t. Using these equations, Minkowski added some brief calculations concerning the orbits and expected revolution times of planets and inferred – using the known values of the mass of the Sun as m^* and of the axis of the Earth's orbit – that his formulas yielded values for the eccentric anomalies in the planetary orbits of the order of 10^{-8}. He concluded with two remarks: first, that the kind of attraction law derived here and the assumption of the postulate of relativity together imply that gravitation propagates with the velocity of light. Second, that considering the small value obtained above for Kepler's equation for eccentric anomalies, the known astronomical data cannot be used to challenge the validity of the laws of motion and modified mechanics proposed here and to support Newtonian mechanics.[45]

Minkowski's treatment of gravitation was extremely sketchy and tentative. An attentive reading of it raises more questions that it seems to answer. Some of these questions have been formulated in the foregoing paragraphs, but more can be added. For instance: Is Minkowski's gravitational force in any sense symmetric with respect to F and F^*? What kind of conservation laws arise within such a theory? Minkowski did not address these issues, either in the article or elsewhere. Rather than addressing the issue of gravitation in detail, when writing this article Minkowski's main concern was clearly to investigate the logical status of the principle of relativity as applied to all physical domains and the plausibility of assuming that it must also hold when dealing with gravitation.

Still, the theory outlined in this lecture was, together with Poincaré's, the starting point of the attempts to extend the validity of the principle of relativity to cover gravitation as well. Einstein himself addressed the same task in an article submitted for publication on December 4, 1907, less than 3 weeks before Minkowskii's talk, in which he raised for the first time the question whether the principle of relativity could be extended to cover accelerated, rather than only inertial reference systems.[46] Although Einstein formulated here for the first time what he later called the principle of equivalence – a fundamental principle of his general theory of relativity – his 1907 attempt did not directly lead to an extension of the validity of relativity. Einstein did not return to this topic until 1911, when his actual efforts to generalize relativity really began. In his 1907 paper Einstein mentioned neither Minkowski nor Poincaré. Nor did Minkowski mention this article of Einstein, and one wonders if at this point he had already read it. Minkowski's approach to electrodynamics and the principle of relativity came to provide the standard language for future investigations, but his specific argumentation on gravitation attracted little if any attention. Minkowski himself mentioned the issue of gravitation once again in his next article, "Space and Time," but only in passing. Arnold Sommerfeld, whose 1910 article contributed more than any other work to systematize and disseminate

[45] [32, p. 404].

[46] [8].

Minkowski's four- and six-vector formalism, claimed that Minkowski's approach to gravitation was no better than Poincaré's, and that if they differed in any respect – as Minkowski had claimed in his article – it was in their methods rather than in their results.[47] Unfortunately, we do not know how Minkowski would have reacted to Sommerfeld's interpretation on this point.

I summarize this section by stressing that Minkowski sought to investigate, in axiomatic terms, the conceptual consequences of applying the postulate of relativity in domains other than electrodynamics. In this framework he addressed, besides mechanics, gravitation and showed how an argument could be worked out for the claim that there was no *prima facie* reason to assume that the postulate of relativity contradicts the observable effects of phenomena pertaining to this latter domain. He concluded that one could envisage the possibility of a truly articulate Lorentz-covarianttheory of gravitation which would approximate the Newtonian theory as a limiting case. It seems, however, that neither Minkowski nor Hilbert considered this theory as anything more than a very preliminary attempt. On the other hand, this whole lecture, and especially its final sections, helps clarifying the kind of motivations underlying Minkowski's investigation of the place of the principle of relativity in physics. Moreover, this particular talk of 1907 shows very clearly how the geometric element ("geometric" taken here in its intuitive-synthetic, rather than in its formal-analytical, sense) entered Minkowski's treatment only gradually, and that an immediate visualization, in geometric terms, of the consequences of the adoption of the principle of relativity in mechanics was not an initial, major motivation behind his attempt. Such a geometrical elements becomes central only in his next text on electrodynamics, "Space and Time".

1.4 Space and Time

Minkowski first presented his views on relativity outside Göttingen on September 21, 1908, when he delivered a lecture at the annual meeting of the German Society of Natural Scientists and Physicians in Köln. The text of his lecture was later published as "*Raum und Zeit*", Minkowski's best known contribution to the special theory of relativity and to the new conception of space and time associated with it. Both the opening and the closing passages of the text have repeatedly been quoted as encapsulating the essence of Minkowski's views. The opening passage of the talk was a rather dramatic proclamation:

> Gentlemen! The conceptions of space and time which I would like to develop before you arise form the soil of experimental physics. Therein lies their strength. Their tendency is radical. Henceforth, space by itself, and time by itself, are doomed to fade away in the shadows, and only a kind of union of the two will preserve an independent reality.[48]

---

[47] [53, p. 687]. On pp. 684–689 one finds a somewhat detailed account of the physical meaning of Minkowski's sketch for a theory of gravitation, and a comparison of it with Poincaré's.

[48] [33, p. 431].

In the closing passage he concluded: "The validity without exception of the world-postulate, I would like to think, is the true nucleus of an electromagnetic image of the world, which, discovered by Lorentz, and further revealed by Einstein, now lies open in the full light of day." These two passages have helped consolidate the image of Minkowski's geometrically motivated approach to relativity and of his alleged commitment to the electromagnetic view of nature. Still, an analysis of his text against the background of Hilbert's program for the axiomatization of physical theories, and in the spirit of the previous two sections, makes clear that such a commitment did not exist, and at the sane allows interpreting these passages in a different way, as will be seen now.

Minkowski started by presenting two kinds of invariance that arise in connection with the equations of Newtonian mechanics. First, the invariance associated with an arbitrary change of position, and second, the one associated with uniform translation. Our choice of a particular point as $t = 0$ does not affect the form of the equations. Although these two kinds of invariance can both be expressed in terms of the groups of invariance they define with respect to the differential equations of mechanics, traditional attitudes towards these respective groups had been utterly different. For, whereas the existence of the group corresponding to the first invariance had usually been seen as expressing a fundamental property of space, the existence of the second (i.e., the group of Galilean transformations) had never attracted any special interest as such. At best, Minkowski said, it had been accepted with disdain (*Verachtung*) in order to be able to make physical sense of the fact that observable phenomena do not enable one to decide whether space, which is assumed to be at rest, is not after all in a state of uniform translation. It is for this reason, Minkowski concluded, that the two groups carry on separate lives with no one thinking to combine them into a single entity.

Minkowski thought that this separation had a counterpart in the way the axiomatic analysis of these two scientific domains had typically been undertaken: in the axiomatization of mechanics, the axioms of geometry are usually taken for granted, and therefore the latter and the former are never analyzed simultaneously, as part of one and the same task.[49] We know precisely what Minkowski meant by this latter assertion. For in his 1905 lectures on the axiomatization of physics, Hilbert had discussed the axiomatization of the laws of motion by adding to the already accepted axioms of geometry separate axioms meant to define time through its two basic properties, namely, its uniform passage and its unidimensionality (*ihr gleichmäßiger Verlauf und ihre Eindimensionalität*).[50] This traditional separation of mechanics and geometry was more explicitly manifest in relation with their respective invariance groups, as explained above, but it had also been implied in the way their axiomatic definitions had been introduced. Minkowski's brilliant idea in this context was to put an end to this separation and to combine the two invariance

[49] [33, p. 431]: "Man ist gewohnt, die Axiome der Geometrie als erledigt anzusehen, wenn man sich reif für die Axiome der Mechanik fühlt, und deshalb werden jene zwei Invarianten wohl selten in einem Atmenzuge genannt." The standard English translation of Minkowski's lecture [35] is somewhat misleading here, as in many other passages.

[50] [16, p. 129]. For details, see [5, pp. 138–153].

groups together. He assumed that this combination would lead to a better understanding of the reality of space and time, and of the laws of physics. The aim of his talk was to explain the implications of such a move.

Minkowski's audience was mainly composed of natural scientists rather than mathematicians. This certainly influenced the kinds of arguments he used and the emphases he chose to adopt. In particular, he stressed from the outset that the ideas presented in the lecture were independent of any particular conception of the ultimate nature of physical phenomena. As in his two previous lectures on the same topic, Minkowski intended his arguments to be an exploration of the logical consequences of adopting the postulate of relativity in the various domains of physics, without necessarily committing himself to any particular view. Therefore, he put forward his arguments in a way intended to prevent any physicist, whatever his basic conception of physical phenomena, from reacting to these ideas with *a priori* suspicion or hostility. Thus, Minkowski's arguments were meant to be compatible with any possible belief concerning the ultimate nature of mass, electromagnetic processes and the ether, and the relationships among these: "In order not to leave a yawning void anywhere," he said, "we want to imagine, that at any place in space at any time something perceptible exists. In order not to say matter or electricity, I will use the word 'substance' to denote this something."[51] *Substance* was therefore a general category rather than being bound to a particular physical interpretation of mass, ether, electricity or any other candidate. In a later passage in which he referred to the velocity of light in empty space, he exercised the same kind of caution: "To avoid speaking either of space or of emptiness, we may define this magnitude in another way, as the ratio of the electromagnetic to the electrostatic unit of electricity."[52]

Assuming that we are able to recognize a substantial point as it moves from a first four-coordinate "world-point," to a second one, Minkowski declared in the introduction that the world can be resolved into world-lines, namely, collections of all the world-points associated with a substantial point when t takes all values between $-\infty$ and ∞. He added that the laws of physics attain their most perfect expression when formulated as relations between such world-lines.

1.4.1 Groups of Transformations

In his first talk on the principle of relativity in 1907, Minkowski had already shown that the assumption of the principle of inertia implies that the velocity of propagation of light in empty space is infinite. This time he discussed this implication, while focusing on certain formal properties of the groups defined by the Galilean transformations and by the Lorentz transformations. The first group expresses the fact that if the x, y, z axes are rotated around the origin of coordinates while $t = 0$, then

[51] [33, p. 432].

[52] [33, p. 434; 35, p. 79].

the expression $x^2 + y^2 + z^2$ remains invariant. The second group expresses the fact that the laws of mechanics remain unchanged under the transformations that send x, y, z, t to $x - \alpha t, y - \beta t, z - \gamma t, t$, with any constant coefficients α, β, γ. Under these transformations, the t-axis can be given whatever upward direction we choose. But how is the demand of orthogonality in space, asked Minkowski, related to this complete freedom of the t-axis? To answer this question Minkowski suggested that one must consider four-dimensional space-time and a more general kind of transformation, namely, those that leave invariant the expression $c^2t^2 - x^2 - y^2 - z^2 = 1$. These transformations turn out to depend on the value of the parameter c and thus classical mechanics appears as a special case of a more general class of theories. He stressed the geometrically intuitive elements of his arguments, by focusing on the case $c^2t^2 - x^2 = 1$, which is graphically represented as a hyperbola on the plane x, t (Fig. 1.3):

Here OB is the asymptote ($ct - x = 0$), and the orthogonal segments OC and OA have the values $OC = 1$ and $OA = 1/c$. Choose now any point A' on the hyperboloid, draw the tangent $A'B'$ to the hyperbola at A', and complete the parallelogram $OA'B'C'$. If OA' and OC' are taken as new axes, x', t' respectively, and we set $OC' = 1, OA' = 1/c$, then the expression for the hyperbola in the new coordinates retains its original form $c^2t'^2 - x'^2 = 1$. Hence, OA' and OC' can now be defined as being themselves orthogonal and thus the hyperbola construction helps to conceive orthogonality in a way that departs from the usual Euclidean intuition. The parameter c determines in this way a family of transformations that, together with the rotations of space-time around the origins of coordinates, form a group, the group G_c. But then – again from geometric considerations – one sees that when c grows infinitely large, the hyperbola approximates the x-axis and, in the limit case, t' can be given any upward direction whatever, while x' approaches x indefinitely. This geometrical argument thus shows that G_∞ is nothing but the above described group of transformations G_c associated with Newtonianmechanics.

This illuminating connection between the two main groups of transformations that arise in physics allowed Minkowski to digress again and comment on the relation between mathematics and physics:

> This being so, and since G_c is mathematically more intelligible than G_∞, it looks as though the thought might have struck some mathematician, fancy-free, that after all, as a matter of fact, natural phenomena do not possess an invariance with the group G_∞, but rather with a

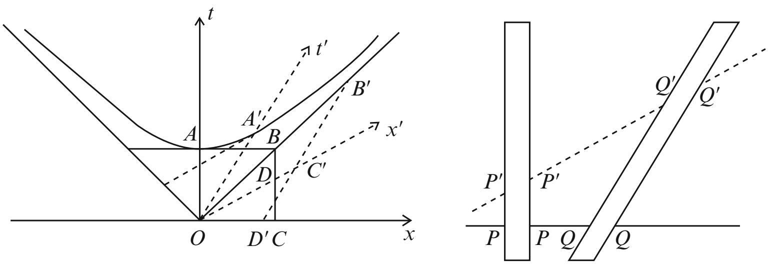


Fig. 1.3 Original diagram of Minkowski's "Space and Time" talk

> group G_c, c being finite and determinate, but in ordinary units of measure, *extremely great*. Such a premonition would have been an extraordinary triumph for pure mathematics. Well, mathematics, though it can now display only staircase-wit, has the satisfaction of being wise after the event, and is able, thanks to its happy antecedents, with its senses sharpened by an unhampered outlook to far horizons, to grasp forthwith the far-reaching consequences of such a metamorphosis of our concept of mature. [33, p. 434; 35, p. 79])

It is not evident, on first reading, what Minkowski meant here when he said that G_c is "mathematically more intelligible" than G_∞, but apparently he was pointing to the fact that the group of Galilean transformations, which in itself had failed to attract any interest from mathematicians, becomes much more mathematically interesting when seen in the more general context of which it appears as a limiting case. In retrospect, Minkowski concluded, this situation might seem to suggest that mathematical insight could have sufficed to realize what is involved here, but in fact this was not the case, and physical considerations were necessary.

The invariance under the group G_c of the laws of physics in a four-dimensional space-time has for Minkowski an additional, important consequence that reinforces – from a different perspective and in a much more compelling fashion – a point of view earlier elaborated in Hilbert's writings, namely, the view of geometry (i.e., the science of sensorial space) as a natural science on which all other physical sciences are grounded. Yet, what Hilbert had initially expressed as an epistemologically grounded conception, and had later developed when discussing the axioms of mechanics on the basis of the axioms of geometry, appears here in the opposite direction: the latest developments of physical science have raised the need to reconsider our basic conception of space and time in such a way as to recognize that geometry is essentially embedded in physics. Thus, to conclude this section of his lecture Minkowski said:

> In correspondence with the figure described above, we may also designate time t', but then must of necessity, in connection therewith, define space by the manifold of the three parameters x', y, z, in which case physical laws would be expressed in exactly the same way by means of x', y, z, t', as by means of x, y, z, t. 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. (ibid.)

1.4.2 Empirical Considerations

So much for the formal, geometrical considerations, but of course the question arises: what empirical facts compel us to adopt this new conception of space? Moreover, can we be sure that this conception never contradicts experience? Is it useful in describing natural phenomena? These questions were discussed by Minkowski in the following three sections of his talk. First, he observed that by means of a suitable transformation the substance associated with a particular world-point could always be conceived as being at rest. This he considered to be a fundamental axiom of his theory of space-time. A direct consequence of the axiom is that every possible velocity in nature is smaller than c. In his second 1907 lecture Minkowski had taken this consequence in itself as a central axiom of the electrodynamics of moving bodies. Formulated in these terms, he felt, it had a somewhat "unpleasant" appearance that raised mistrust, but in the present four-dimensional formulation it could be grasped more easily.

Using the groups G_c and G_∞, Minkowski explained the problems raised by the Michelson experiment, given the different invariance groups characteristic of different physical disciplines. He stressed that the concept of a rigid body may have a coherent meaning only in a mechanics based on the group G_∞, and that the contraction hypothesis had been introduced by Lorentz in order to account for the divergence detected between theory and experiment. Remarkably enough, in spite of having stressed pompously in the opening passage of his talk that the origin of these new conceptions was fully rooted in experiment, this is the only reference in the whole text to anything of the sort. In fact, Minkowski preferred to ignore recent results by Kaufmann already mentioned above, that allegedly refuted the theory of relativity.[53] Admitting that the contraction hypothesis in its original form "sounds extremely fantastical," he proceeded to show that it is entirely coherent when seen in terms of] the new conception of space and time, and that the latter clarified the former completely. Minkowski's explanation was fully geometrical and it relied on a straightforward verification of the properties of a rectangle and a parallelogram drawn on the two-dimensional figure introduced in the first section. At this point Minkowski also characterized Einstein's contribution in this context, as explaining the nature of local time. Whereas Lorentz had introduced the concept as a tool for better understanding the contraction hypothesis, Einstein "clearly recognized that the time of the one electron is just as good as that of the other."[54] Thus, Minkowski saw that Einstein had essentially undermined the idea of *time* as a concept unequivocally determined by phenomena. But then, in spite of the importance of this achievement, neither Einstein nor Lorentz undertook a similar attack on the concept of *space*. Minkowski considered such an attack to be indispensable in uncovering the full implications of the postulate of relativity, and he saw his own ideas as having contributed to the full achievement of that aim. It was in this framework that he introduced the term "World-postulate" instead of relativity:

> When [the attack on the traditional concept of space] has been undertaken, the word *relativity-postulate* for the requirement of invariance with the group G_c seems to me very feeble. Since the postulate comes to mean that spatio-temporal phenomena manifest themselves only in terms of the four-dimensional world, but the projection in space and in time may still be performed with certain liberty, I prefer to call it the *postulate of the absolute world* (or briefly, the world-postulate). [33, p. 437][55]

[53] This point has been raised by Scott Walter [56, p. 52] in his perceptive study of the rhetoric strategy followed by Minkowski, the mathematician, in addressing a public of non-mathematicians.

[54] [33, p. 437 ([35, p. 83])]. In his obituary of Minkowski, Hilbert [17, p. 90] repeated this assessment. For a discussion of the differences in the conception of time in Einstein's and in Minkowski's theories, see [56], § 3.5.

[55] Minkowski's original sentence – "noch mit einer gewissen Freiheit vorgenommen werden kann, ..." – appears in the English translation [35, p. 83] as: "may still be undertaken with a certain degree of freedom." This rendering seems to me somewhat misleading in this context.

It is significant that in this talk Einstein's work becomes a much more important focus of reference for Minkowski than in the previous two, particularly Einstein's innovative conception of time. It is very likely that by this time Minkowski had already read Einstein's 1907 article mentioned above. This survey article had been written at the request of Johannes Stark (1874–1957), editor of the *Jahrbuch der Radioaktivität und Elektronik*, following the recent publication of Kaufmann's criticism of relativity. Attempting to strengthen the theoretical and experimental support for his theory, Einstein now stressed the *similarities* between Lorentz's and his own work. He presented the latter as genetically related to the former (and, implicitly, also superior to it) rather than presenting these as two alternative approaches to the same problem. At the same time he explicitly attributed a central place to the Michelson-Morley experiment in the development of the whole theory (and implicitly in the development of his own).[56] Einstein himself considered this presentation of his theory to be simpler and more intuitive than the one of 1905 where he had striven, above all, for "unity of presentation".[57] The rhetoric of Minkowski's talk connects smoothly and in visible ways with the spirit and contents of Einstein's 1907 article.

In the third part of the lecture, Minkowski showed that the world-postulate provides a much clearer understanding of the laws of physics, by allowing a symmetrical treatment of the four coordinates x, y, z, t. In this first section he introduced the concept – only implicit in his earlier lectures – of a light-cone (in fact, he only spoke separately of the front- and back-cones of a point O) and explored its usefulness, especially in dealing with the concept of acceleration.

1.4.3 Relativity and Existing Physical Theories

In the last two sections, Minkowski addressed again the main point discussed in his previous talk, namely, the compatibility of the principle of relativity with existing physical theories, or, as he put it here, that "the assumption of the group G_c for the laws of physics never leads to a contradiction." In order to show this, Minkowski understood that it was "unavoidable to undertake a revision of the whole of physics on the basis of this assumption." Such a revision had in fact already begun. Minkowski cited again Planck's recent article on thermodynamics and heat radiation,[58] as well as his own earlier lecture, already published by then, where the compatibility of the postulate of relativity with the equations of electrodynamics and of mechanics (retaining, he stressed, the concept of mass) had been addressed. With reference to the latter domain, Minkowski elaborated this time on the question of how the

[56] Cf. [55, pp. 275–281]. For debates on the actual role of the Michelson-Morley in the development of Einstein's ideas and its historiography, see [15; 18, pp. 279–370; 54]. At any rate, Einstein had read about the experiment as early as 1899. Cf. *CPAE* 1, Doc. 45, 216.

[57] Einstein to Stark, November 1, 1907 (*CPAE* 5, Doc. 63).

[58] [40]. Another remarkable aspect in the rhetoric of Minkowski in this talk is the total absence of references to Poincaré. On possible reasons for this, see [56, pp. 60–62].

expressions of force and energy change when the frame of reference changes. He then showed how the effects produced by a moving point-charge, and in particular the expression of its ponderomotive force, can be best understood in terms of the world postulate. He stressed the simplicity of his own formulation as compared with what he considered the cumbersome appearance of previous ones.

Finally, in a brief passage, Minkowski addressed the question of gravitation, noting that the adoption of the world-postulate for mechanics as well as for electrodynamics eliminated the "disturbing lack of harmony" between these two domains. Referring back to his published lecture of 1907, he asserted that, by introducing in the equations of motion under gravitation the proper time of one of the two attracting bodies (which is assumed to be moving, while the other is at rest), one would obtain a very good approximation to Kepler's laws. From this he concluded, once again, that it is possible to reformulate gravitation so as to comply with the world-postulate.

In his closing remarks, Minkowski addressed the question of the electromagnetic world-view and the postulate of relativity, which he had expressly bypassed throughout the lecture. For Minkowski, it was not the case that all these physical domains were compatible with the world-postulate (merely) because their equations had been derived in a particular way; the postulate had a much more general validity than that. It is in this light that we must understand the often-quoted closing passage of the lecture. The equations that describe electromagnetic processes in ponderable bodies completely comply with the world-postulate, Minkowski remarked. Moreover, as he intended to show on a different occasion, in order to verify this fact it is not even necessary to abandon Lorentz's erudite (*gelehrte*) derivation of these fundamental equations, based on the basic conceptions (*Vorstellungen*) of the theory of the electron.[59] In other words, whatever the ultimate nature of physical processes may be, the world-postulate, i.e., the universal demand for invariance under the group G_c of the equations expressing the laws of physical processes, must hold valid. This is what we have learnt from the latest developments in physics and this is what Minkowski expressed in his well-known assertion:

> The validity without exception of the world-postulate, I like to think, is the true nucleus of an electromagnetic image of the world, which, discovered by Lorentz, and further revealed by Einstein, now lies open in the full light of day. In the development of its mathematical consequences there will be ample suggestions for experimental verification of the postulate, which will suffice to conciliate even those to whom the abandonment of the old-established views is unsympathetic or painful, by the idea of a pre-established harmony between mathematics and physics [33, p. 444; 35, p. 91].

Clearly, then, in reading this passage we need not assume that Minkowski was trying to advance the view that all physical phenomena, and in particular the inertial properties of mass, can be reduced to electromagnetic phenomena. Nor is it necessary to determine to what extent Minkowski had understood Einstein's innovative point of view in his paper on the electrodynamics of moving bodies, as compared to all

[59] [33, p. 444]. Also here the translation [35, pp. 90–91] fails to convey the meaning of the original passage.

the other sources from which his theory took inspiration. Rather, Minkowski only claimed here that the electromagnetic world-view is nothing but what the world-postulate asserts: the belief in the general validity of the world-postulate is all that there is, and can be, to the electromagnetic world-view. A similar attitude was found in Hilbert's 1905 lectures on physics, when he analyzed in axiomatic terms the basic assumptions of a theory that are necessary for the derivation of its main theorems, but avoided, as much as possible, any commitment to a particular world-view. Both Minkowski and Hilbert believed that in constructing the mathematical skeleton of all physical theories, certain universal principles must be postulated (the world-postulate and general covariance, but also the energy principle and the continuity principle); even in the face of new empirical discoveries that will force changes in the details of individual theories, these general principles will continue to hold true. Moreover, the idea of a pre-established harmony of mathematics and physics, so popular in the discourse of the Göttingen scientific community, can be traced back to the belief in the existence of such universal principles, rather than to the specific contents of particular, probably provisional, physical theories expressed in mathematical terms. The idea of a "true nucleus" (*der wahre Kern*) of physical theories that is preserved amidst other, presumably more cosmetic traits, will also resurface in remarkable circumstances in the work of Hilbert on general relativity.[60]

1.5 Max Born, Relativity, and the Theories of the Electron

In "Time and Space", Minkowski had set to verify the universal validity of the postulate of relativity at the macroscopic level. In the closing passages of the lecture he declared that on a future occasion he intended to do so at the microscopic level as well, namely, starting from Lorentz's equations for the motion of the electron. On July 28, 1908, he gave a talk at the meeting of the Göttingen Mathematical Society on the basic equations of electrodynamics. Although no complete manuscript of this lecture is known, a very short account, published in the *JDMV* seems to indicate that Minkowski addressed precisely the microscopic derivation of the equations using the principle of relativity.[61] Be that as it may, he was not able to publish any of these ideas before his untimely death on January 12, 1909. We nevertheless have a fair idea of what these ideas were, from an article published by Max Born in 1910, explicitly giving credit for its contents to Minkowski. Born used Minkowski's unfinished manuscripts and the ideas he heard in the intense conversations held between the two before Minkowski's death.

According to Born's introduction, the starting point of Minkowski's "*Grundgleichungen*" had been the assumption of the validity of the Maxwell equations for stationary bodies, inductively inferred from experience. This point of view,

[60] See [5, pp. 399–403].

[61] *JDMV* 17 (1909), 111.

explained Born, differed from Lorentz's, which accounted for processes in material bodies in terms of certain hypotheses about the behavior of the electrons that compose those bodies. Lorentz had considered three kinds of electrons. First, there were conduction electrons (*Leitungselektron*), whose movement is independent of matter and whose charge constitutes "true electricity." Second, polarization electrons provided a state of equilibrium inside molecules of matter; these electrons, however, can be dislocated from this state through the action of the electromagnetic field. The variable electricity density produced in this way is known as the "free electricity." Third came the magnetization electrons that orbited around central points inside matter, thus giving rise to magnetic phenomena. Lorentz's equations for electromagnetic processes in material bodies were based on the mean values of the magnitudes of the convection current due to the three types of electron. Yet as Minkowski had shown in his "*Grundgleichungen*", in certain cases – specifically, in the case of magnetized matter – the equations thus obtained contradict the postulate of relativity.

The specific aim of the article, then, was to extend the validity of the postulate to cover all cases, including the problematic one pointed out by Minkowski in his earlier article. But for all the assumptions concerning the complex structure of matter that the above discussion implies, Born understood the need to stress, as Minkowski had done before him, the *independence* of this study from a particular conception of the ultimate nature of matter, ether or electricity. He thus explained that "among the characteristic hypotheses of the electron theory, the atomic structure of electricity plays only a limited role in Lorentz's derivation of the equations," given the fact that mean values have been taken over "infinitely small physical domains", so that all this structure is completely blurred, and the mean values, in the final account, appear as continuous functions of time and location. Born thus justified his adoption of Lorentz's approach to the derivation of the equations, without thereby committing himself to any ontological assumptions. He declared very explicitly:

> We hence altogether forgo an understanding of the fine structure of electricity. From among Lorentz's conceptions, we adopt only the assumptions *that electricity is a continuum that pervades all matter, that the former partially moves freely inside the latter and partially is tied to it, being able to carry out only very reduced motions relative to it.*

If we want to come as close as possible to Lorentz, then all the magnitudes introduced below should be considered as Lorentzian mean values. It is however not necessary to differentiate among them, using special symbols, as if they were related to the various kinds of electrons, since we never make use of the latter.[62]

Following Minkowski's death, Born went on to develop his own ideas on relativity, which he had begun to consider following his reading of Einstein. A fundamental contribution of Born was the introduction of the Lorentz-invariant concept of a rigid body, a concept to which Born was led while working on the problem of the self-energy of the electron. As we saw above, Minkowski had already made it clear in

[62] Minkowski 1910, 61 (Italics in the original).

"Space and Time" that the traditional concept of rigid body did not make sense outside Newtonian mechanics. Born's interest in this question implied an involvement in the Abraham-Lorentz debate concerning the independence or dependence of the mass of the (rigid or deformable) electron on its velocity, and, in the question of the possible electromagnetic nature of the mass of the electron. In his autobiography, Born mentions that in their discussions of these issues, Minkowski "had not been enthusiastic about Born's own ideas but had raised no objections."[63] One wonders whether Minkowski's lack of enthusiasm was not perhaps connected to Born's particular interest in the electromagnetic mass of the electron, a topic which Minkowski persistently tried to avoid in his own work.

Both Abraham and Lorentz had calculated the self-energy of a charged, rigid body moving uniformly and used this energy as the Hamiltonian function for deriving the equations of motion. Born doubted the validity of an additional assumption implicit in their calculations, namely, that the energy calculated for uniform motion is the same for accelerated motion, since in an accelerated body different points have different velocities and therefore, according to the principle of relativity, different contractions. The classical concept of a rigid body is thus no longer applicable. Without entering to all the technical details of Born's derivation, I will nevertheless mention that his definition is based on finding a Lorentz-covariant expression of the distance between any two space-time points; the classical distance between two points in a body is given by

$$r_{ij}^2 = (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2,$$

which is clearly not Lorentz-covariant.[64]

Born discussed the Lorentz-covariant definition of rigidity in two articles published in 1909. In the first, submitted on January 9 (just 3 days before Minkowski's death), he discussed the relation between the concept of mass and the principle of relativity. This article still reflects the direct influence of Minkowski's point of view. Born referred in the introduction of this article to the "Abraham-Sommerfeld theory of the rigid electron", whose main task he described as that of reducing the inertial mass of the electron to purely electrodynamic processes. The theory, however, does not satisfy the "Lorentz-Einstein principle of relativity." On the other hand, said Born, the latter principle has not led to a satisfactory explanation of inertial mass. The equations of motion formulated by Lorentz, Einstein and Minkowski are suggestive approximations of the Newtonian equations, which at the same time satisfy the relativity principle of electrodynamics. The concept of mass is thus modified in the works of the three so as to fit that principle without, however, explaining the concept in electrodynamical terms.

Born's treatment of mass was intended as an analogy to Minkowski's ideas, but applied in the framework of the Abraham-Sommerfeld theory. Minkowski had

---

[63] [3, p. 132].

[64] For a more detailed discussion of Born's concept of rigid body and its impact, see [26; 27, pp. 243–257].

modified the Hamiltonian principle of classical mechanics so as to make the ensuing equations of motion fit the relativity principle. The variational equation to which this principle gives raise yields two integrals, one of which expresses the effect of the mass. Born intended to introduce a similar generalized Hamiltonian involving only electromagnetic magnitudes, and to derive the mass in a way similar to Minkowski's. However, it is noteworthy that for all of his interest in the Abraham-Sommerfeld theory, Born took pains to stress explicitly that his derivation was in no way dependent on any assumption concerning the ultimate nature of electricity – in particular, those that underlie Abraham's and Lorentz's theories. Clearly alluding to the point of view adopted in the paper he had published under Minkowski's name, Born wrote:

> *It must be emphasized that no use will be made here of atomistic hypotheses*. In fact, the atom or the electron, imagined as rigid bodies, can in no way be incorporated into the system of electrodynamics built on the principle of relativity, in which no analog is known of a rigid body in arbitrary accelerated motion. However, given the fact that all the basic expressions of Lorentz's theory of the electron seem to be independent of the hypotheses concerning the atomistic electron, the inertia of a continuously flowing charge can be likewise electromagnetically established in the sense suggested above. Naturally, this conception in no way contradicts those physical facts that indicate an extraordinarily strong, variable (almost atomistic) spatial distribution of matter and electricity.[65]

Born's second publication that year on the same topic is his better-known paper containing the definition of rigid bodies, submitted on June 13. Born asserted that his definition of rigidity would play a role in Maxwellian electrodynamics similar to that played by the classical rigid body in Newtonian mechanics. He was now ready to express opinions on fundamental issues openly, yet he preserved much of Minkowski's characteristic caution. His theory, he thought, accounted for the atomistic structure of electricity in a way that Abraham's theory did not. It thus corresponded to the "atomistic instinct" of so many experimentalists who found it very hard to support recent attempts to describe the movement of electricity as a fluid, unconstrained by any kinematic conditions, and affected only by the action of its own field.[66] On the other hand, in motivating this analysis Born did invoke concerns like those repeatedly stressed by Minkowski: to allow for a further clarification of the conceptual relationship between electrodynamics and the principle of relativity. This view, which is manifest in various places in Born's paper, is best encapsulated in the following passage:

> *The practical value of the new definition of rigidity must manifest itself in the dynamics of the electron*. The greater or lesser transparency of the results obtained by means of it will also be used, to a certain extent, for or against making the assumption of the principle of relativity universally valid, since experiments have not yet provided a definite proof of it and perhaps never will.[67]

[65] [1, pp. 572–573] (Italics in the original).

[66] [2, pp. 5–6].

[67] [2, p. 4] (Italics in the original).

1.6 Summary and Concluding Remarks

In this chapter I have argued that in order to understand the proper historical context of Minkowski's work on relativity one must consider it against the background of the ideas that animated Hilbert's program for the axiomatization of physics. In turn, Minkowski's work clarify the potential scope and possible applications of the principles of Hilbert's program, albeit in a direction that Hilbert did not cover – and could not have imagined – when he formulated the sixth problem of his Paris address in 1900 (a call for the axiomatizaion of physics) and even in teaching his 1905 course in Göttingen.

The assumption of universal validity of Lorentz covariance had been strongly suggested by experimental results obtained during the late nineteenth century, and its theoretical implications had been investigated from different perspectives in recent works, noticeably those of Lorentz, Poincaré and Einstein. Yet, in a spirit similar to that underlying Hilbert's program, Minkowski believed that the logical structure of the physical theories built on the principle of relativity had not been satisfactorily elucidated, and he set out to do so. He was interested in exploring the logical consequences of the principle and in proving that it does not contradict the existing edifice of the various disciplines of physics. The postulate of relativity should be taken as a further axiom appearing at the base of each and every physical theory, together with the particular axioms of that theory. Minkowski was able to prove for certain domains of physics that the ensuing theory indeed produced a consistent logical structure. For some other theories, such as gravitation, he was less successful, but he claimed to have showed at least that no contradiction had arisen by adding the principle, and that a consistent, Lorentz-covariant theory of gravitation could eventually be worked out in detail.

But the postulate of relativity was for Minkowski not simply an additional axiom, with perhaps a wider domain of validity in physics than others. It was an axiom of a different nature: a principle that should be valid for every conceivable physical theory, even those theories that were yet to be discovered or formulated. Minkowski compared the status of the postulate of relativity with that of the principle of conservation of energy, whose validity we assume even for yet unknown forms of energy. Interestingly, Einstein, too, had drawn a similar comparison at roughly the same time, between the principle of relativity and the second law of thermodynamics. Minkowski may have been aware of this, since it appeared in the *Annalen der Physik* as a reply to an earlier article of Paul Ehrenfest (1880–1933), who was then at Göttingen. But Einstein and Minkowski compared relativity and conservation of energy in different ways. Einstein spoke in his article of two "open" principles of physics, with a strong heuristic character. Unlike Minkowski and Hilbert, Einstein did not see the principle of relativity and the principle of energy conservation as parts of strictly deductive systems from which the particular laws of a given domain could be derived.[68] More generally, although Einstein introduced the principle of

[68] [8].

relativity together with the constancy of light at the beginning of his 1905 article as "postulates" of the theory (in some sense of the word), there are clear differences between Einstein's approach and Minkowski's axiomatic analysis of the postulate of relativity.[69] In fact, one of the main aims of Hilbert's program was to address situations like that raised by Einstein, which he saw as potentially problematic. As Hertz had pointed out in the introduction to his *Principles of Mechanics*, it has often been the case in the history of physics that, faced with conflict between an existing theory and new empirical findings, physicists have added new hypotheses that apparently resolve the disagreement but perhaps contradict some other consequences of the existing theory. Hilbert thought that an adequate axiomatic analysis of the principles of a given theory would help to clear away possible contradictions and superfluities created by the gradual introduction of new hypotheses into existing theories. This was essentially the same goal pursued by Minkowski: he sought to verify that the recent introduction of the principle of relativity into physics had not created such a problematic situation.

One of the central points that emerges from studying Minkowski's work within its proper context, and one which is strongly suggested by the proximity of Hilbert's program, is the idea that the place of the postulate of relativity in physics could be fully analyzed without assuming, and certainly without committing oneself to, any particular conception of the ultimate nature of physical phenomena. We may assume that, to the extent that he did take a definite position on the foundations of physics, he must have been close to some kind of mechanical reductionism, similar to that of Hilbert at the time. While there seems to be no direct evidence to answer this question, Minkowski's admiration for Hertz was consistently expressed and there is no evidence showing that he opposed him on this particular point.

The axiomatizing motivation behind Minkowski's work provides, then, a main perspective from which to understand the roots and the goals of his overall involvement with electrodynamics and relativity. This kind of motivation, however, appeared in combination with several other elements that informed his much more complex mathematical and physical background. The geometric element of this background, for instance, is one that has received much attention in the secondary literature, and must certainly be taken into account. Still, there are several reasons why one should be cautious in assessing its actual significance. For one, the very terms "geometry" and "geometrical" are much too comprehensive and sometimes imprecise. They need to be sharpened and placed in proper historical context if they are to explain in some sense Minkowski's motivations or the thrust of his articles on electrodynamics.[70] One should be able to describe, for instance, Minkowski's views on some of the basic, foundational questions of geometry and mathematics in general. We do not have much written evidence of this, besides the few

[69] On the other hand, Minkowski's axiomatic approach, and in particular his stress on universally valid principles in physics, strongly brings to mind Einstein's oft-quoted remarks on the differences between theories of principle and constructive theories. Cf. *CPAE* 2, *xxi–xxii*.

[70] A convincing analysis of the role of geometrical visualization in Minkowski's work in number theory appears in [51].

statements quoted at the beginning of this chapter that indicate a proximity to Hilbert's empiricist inclinations, and a stress on the significant, potential contributions of physical ideas to pure mathematics.

Elucidating the specific nature of Minkowski's conception of geometry becomes particularly important if we are to understand why, once he decided to undertake the axiomatic clarification of the role of the principle of relativity in physics, Minkowski came forward with a space-time geometry as an essential part of his analysis. Of primary interest in any discussion of this issue must be the connection between groups of transformations and geometry, which in "Space and Time", as was seen above, becomes a focal point of Minkowski's analysis. Klein was evidently very excited about this particular feature, and in a lecture of May 1910 he suggested, while referring to work done back in 1871, that he had in fact anticipated the approach behind Minkowski's study of the Lorentz group. The Minkowski space, he suggested, was just the four-dimensional version of a mathematical idea long familiar to himself, as well as to geometers like Sophus Lie (1842–1899) or Gaston Darboux (1842–1917).[71] On the other hand, when lecturing in 1917 on the history of mathematics in the nineteenth century, Klein remarked that among Minkowski's four papers he liked the first one most. Klein stressed the invariant-theoretic spirit of this paper as the faithful manifestation of Minkowski's way of thought.[72] Minkowski, for his part, did not mention Klein's ideas at all in his own articles, at least not explicitly. One may only wonder what would have been his reaction to Klein's assessments, had he lived to read them.[73] Although the connections suggested by Klein between his early geometrical work and the group-theoretical aspects of relativity in Minkowski's work may seem in retrospect clearly visible, there is no direct evidence that Minkowski was thinking literally in those terms when elaborating his own ideas on space and time.[74] Of course, the general idea that geometries can be characterized in terms of their groups of motions was by then widely accepted, and was certainly part and parcel of Hilbert's and Minkowski's most basic mathematical conceptions. An yet, one remarkable point that comes forward in my presentation is that, in the end, it was based on physical, rather than on purely mathematical considerations, that

[71] Klein expressed these views in a meeting of the GMG, and they were published as [22].

[72] [23, pp. 74–75], referring to [34]. Klein contrasted this paper with the *Grundgleichungen* in which – in order not to demand previous mathematical knowledge from his audience – Minkowski had adopted a more concise, but somewhat ad-hoc, matricial approach. The latter, Klein thought, was perhaps more technically accessible, but also less appropriate for expressing the essence of Minkowski's thoughts.

[73] As already pointed out, the impact of some of some of Klein's work, particularly of the *Erlangen Programm* was somewhat overstated in many retrospective historical analyses, including those of Klein himself. See above §1.2, especially note 78.

[74] For a discussion on the connection between Minkowski's space-time and the ideas associated with Klein's *Erlanger Programm* see [37, p.797]. Norton raises an important point when he claims that "the notion of spacetime was introduced into physics almost as a perfunctory by-product of the *Erlangen* program," but as indicated here, this formulation would seem to imply that program subsumed all the contemporary work on the relations between geometry and groups of transformations, an assumption that needs to be carefully qualified.

Minkowski's work helped consolidate the view that geometry is best understood in terms of the theory of groups of transformations.

The first to establish the explicit connection between the terminology and the ideas of group theory and the Lorentz covariance of the equations of electrodynamics was Poincaré, in his 1905 article. Remarkably, he had also been the first to use four-dimensional coordinates in connection with electrodynamics and the principle of relativity. Minkowski, on the other hand, was the first to combine all these elements into the new conception of the four-dimensional manifold of space-time, a conception that, however, emerged fully-fledged only in his 1908 Köln lecture and was absent from his earlier ones. What was the background against which Minkowski was led to take a step beyond the point that Poincaré had reached in his own work, and thus to introduce the idea of space-time as the underlying concept that embodies the new conception of physics? It is perhaps at this particular point that the specific impact of Einstein's work on Minkowski may have been decisive. One aspect of this work that Minkowski specifically singled out for its importance was Einstein's contribution to modifying the traditional concept of time; Minkowski proposed to do something similar for the concept of space, by replacing it with a four-dimensional geometry of space-time. A combination of this essential point taken from the original work of Einstein, together with the axiomatic perspective stemming from Hilbert's program may have provided the fundamental trigger leading to this innovation. Indeed, when explaining his motivation for studying kinematics with group theoretical tools, Minkowski asserted that the separation between kinematics and geometry had traditionally been assumed both in existing axiomatic analyses and in group-theoretical investigations. Hilbert had explicitly stressed in his axiomatization lectures that the axioms of kinematics would be obtained by coupling to the axioms of geometry, accounting for space, those required in order to account for the properties of time.

The subsequent development of the theory of relativity can hardly be told without referring to the enormous influence of Minkowski's contributions.[75] After an initial stage of indecision and critical responses, the space-time manifold as well as the four-vector language eventually became inseparable from the fundamental ideas introduced by Lorentz, Poincaré, and Einstein. Among the first to insist upon the importance of Minkowski's formulation were Max von Laue (1879–1960) and Sommerfeld. Sommerfeld, who had actually been among the earlier critics of Einstein's relativity, published two articles in 1910 that elaborated in a systematic fashion the ideas introduced by Minkowski and became the standard point reference for physicist over the coming years.[76] Laue published in 1911 the first introductory textbook on the special theory of relativity[77] that precisely because his use of Minkowski's formulation presented the theory in a level of clarity and

[75] For an account of the immediate, varying responses among mathematicians and physicists, see [56], § 4.

[76] [53].

[77] [24].

sophistication that surpassed by far Einstein's original one. Einstein's initial reaction to Minkowski's work, was less enthusiastic, but he soon changed his attitude, and perhaps the influence of Laue and Sommerfeld may have been crucial in this respect.[78]

On the other hand, Minkowski's term "world-postulate", and the connotations implied by it, was never enthusiastically adopted,[79] and even less so was the kind of axiomatic analysis he performed for ensuring that the adoption of the world-postulate at the basis of any branch of physics would not lead to contradiction with the existing theories. And paramount among the existing theories for which the status of relativity remained unclear was gravitation. Physicists did not accord any special attention to Minkowski's more specific axiomatic treatment of the equations of electrodynamics for moving matter either. Hilbert, as usual, followed his own idiosyncratic path, and over the years following Minkowski's death he continued to insist in his lectures upon the need for an axiomatic treatment of physical theories, and to stress the importance of Minkowski's contribution in this regard. Eventually, when in 1915 Hilbert dedicated efforts to finding generally covariant field-equations

[78] In existing accounts, Einstein's alleged negative attitude towards Minkowski's work has sometimes been overemphasized. Thus, for instance, it has been repeatedly said that Einstein considered Minkowski's reformulation of his theory to be no more than "superfluous erudition" The source for this statement is [38, p. 151]. Pais, however, quotes no direct evidence, but rather attributes the claim to Valentin Bargmann (1908–1989), who reportedly heard it from Einstein. Bargmann, it must be emphasized, met Einstein for the first time in 1937. A second, oft-quoted statement in this direction attributes to Einstein the complaint that "since the mathematicians pounced on the relativity theory I no longer understand myself." Such a statement appears in [52, p. 46]. Einstein was also quoted as claiming that he could "hardly understand" Laue's book because of its strongly mathematical orientation, that followed very closely Minkowski's approach (cf. the introduction to the journal *Historical Studies of Physical Science* (*HSPS*) Vol. 7, *xxvii*, quoting [11, p. 206]. Einstein himself wrote in 1942 the preface of the German edition of Frank's book). Frank describes Einstein's claim (which is undocumented, in any case) as having been said "jokingly". The *HSPS* introduction already says "half-jokingly".

Written, relevant evidence that is available leads to different kind of emphases when describing Einstein's attitude in this regard. Thus for instance, Einstein and Laub [9, 10] do avoid the use of four-vectors and claim that Minkowski's mathematics is very difficult *for the reader*. Probably they did not favor Minkowski's formal approach at this stage, but they do not explicitly dismiss it either. In an unpublished article on the Special Theory of Relativity (STR) written in 1911 (*CPAE* 4, Doc.1), Einstein redid much of what appears in his collaboration with Jakob Laub (1882–1962), but now in four-dimensional notation. In fact, already in the summer of 1910, in a letter to Sommerfeld (*CPAE* 5, Doc. 211), Einstein explicitly expressed his increasing appreciation for the importance of such an approach. Cf. also a lecture of Jan. 16, 1991 – Einstein 1911. Whereas in January 1916, in a letter to Michele Besso (*CPAE* 8, Doc. 178), Einstein repeated that Minkowski's papers are "needlessly complicated", he could certainly have recommended a simpler and more elegant presentation in Laue's book.

As for Laue, Einstein consistently praised the high quality and clarity of his book. Cf. e.g., Einstein to Kleiner, April 3, 1912 (*CPAE* 5, Doc. 381). Moreover, in a manuscript written in 1912–1913, and published only recently (*CPAE* 4, Doc. 1, esp. §§3, 4), Einstein presents STR while following very closely the approaches of both Minkowski and Laue.

[79] Indeed, even Born, who was among the first to propagate Minkowski's formalism, did never come to use the term. Cf. [55, p. 293], footnote 67.

of gravitation, he certainly saw himself as following in the footsteps of Minkowski's earlier work, not so much regarding the specific way the latter had attempted to formulate a Lorentz-covarianttheory of gravitation, but rather concerning the principles on which this attempt had been based. Still, the way from Minkowski's treatment of gravitation in 1908–1909 to Hilbert's treatment of the same matter in 1915 was anything but straightforward.

References

1. Born, M.: Die träge Masse und das Relativitätsprinzip. Ann. Phys. **28**, 571–584 (1909)
2. Born, M.: Die Theorie des starren Elektrons in der Kinematik des Relativitätsprinzip. Ann. Phys. **30**, 1–56 (1909a)
3. Born, M.: My Life: Recollections of a Nobel Laureate. Scribner's, New York (1978)
4. Cohn, E.: Über die Gleichungen des electromagnetischen Feldes für bewegte Körper. Ann. Phys. **7**, 29–56 (1902)
5. Corry, L.: Hilbert and the Axiomatization of Physics (1898–1918): From 'Grundlagen der Geometrie' to 'Grundlagen der Physik'. Kluwer, Dordrecht (2004)
6. Darrigol, O.: The electrodynamic revolution in Germany as documented by early German expositions of 'Maxwell's Theory'. Arch. Hist. Ex. Sci. **45**, 189–280 (1993)
7. Darrigol, O.: Electrodynamics form Ampère to Einstein. The University of Chicago Press, Chicago, IL (2000)
8. Einstein, A.: Bemerkungen zu der Notiz von Hrn. Paul Ehrenfest: 'Die Translation deformierbarer Elektronen und der Flächensatz'. Ann. Phys. **23**, 206–208 (1907) (*CPAE* 2, Doc. 44.)
9. Einstein, A., Laub, K.: Über die elektromagnetischen Grundgleichungen für bewegte Körpern. Ann. Phys. **26**, 532–540 (1908) (*CPAE* 2, Doc. 51.)
10. Einstein, A., Laub, K.: Über die im elektromagnetischen Felde auf ruhende Körper ausgeübten ponderomotorischen Kräfte. Ann. Phys. **26**, 541–550 (1908a) (*CPAE* 2, Doc. 52.)
11. Frank, P.: Einstein. His Life and Times. New York, A. A. Knopf (1947)
12. Galison, P.: Minkowski's space-time: From visual thinking to the absolute world. Hist. Stud. Phys. Sci. **10**, 85–121 (1979)
13. Giannetto, E.: The rise of special relativity: Henri Poincaré's work before Einstein's. In: Tucci, P. (ed.) Atti del XVIII Congresso Nazionale di Storia della Fisica e dell'Astronomia (Como 15–16 maggio 1998), pp. 171–207. Milano (1999)
14. Heilbron, J.: The Dilemmas of an Upright Man. Max Planck and the Fortunes of German Science. Harvard University Press, Cambridge, MA (2000)
15. Hentschel, K.: Interpretationen und Fehlinterpretationen der speziellen und der allgemeinen Relativitätstheorie durch Zeitgenossen Albert Einsteins. Birkhäuser, Basel/Boston (1990)
16. Hilbert, D.: Logische Principien des mathematischen Denkens (1905) (David Hilbert Nachlass, Göttingen 558a. Annotated by Max Born).
17. Hilbert, D.: Hermann Minkowski. Gött. Nach. 72–101 (1909) (Repr. in Math. Ann. **68**, 445–471 (1910))
18. Holton, G.: Origins of Scientific Thought: Kepler to Einstein. Harvard University Press, Cambridge, MA (1988)
19. Jungnickel, C., McCormmach, R.: Intellectual Mastery of Nature – Theoretical Physics form Ohm to Einstein, 2 Vols. Chicago University Press, Chicago, IL (1986)
20. Katzir, S.: Poincaré's relativistic Physics and its origins. Phys. Perspect. **7**, 268–292 (2005)
21. Kaufmann, W.: Über die Konstitution des Elektrons. Ann. Phys. **19**, 487–553 (1906)
22. Klein, F.: Über die geometrischen Grundlagen der Lorentzgruppe. Jahresb. DMV **19**, 281–300 (1910)

23. Klein, F.: Vorlesungen über die Entwicklung der Mathematik im 19. In: Courant, R., Neugebauer, O. (eds.) Jahrhundert, 2 Vols. Springer, Berlin (1926) (Chelsea Repr., New York, 1948.)
24. Laue, M. v.: Das Relativitätsprinzip. Braunschweig, Vieweg (1911)
25. Lorentz, H.A.: Weiterbildung der Maxwellschen Theorie. Elektronentheorie. Encyklopädie der mathematischen Wissenschaften mit Einschluss ihrer Anwendungen V, 2–14, 145–280 (1904)
26. Maltese, G., Orlando, L.: The definition of rigidity in the special theory of relativity and the genesis of the general theory of relativity. Stud. Hist. Phil. Mod. Phys. **26B**, 263–306 (1995)
27. Miller, A.I.: Albert Einstein's Special Theory of Relativity: Emergence (1905) and Early Interpretation, (1905–1911). Springer, New York (1997)
28. Minkowski, H.: (*GA*) *Gesammelte Abhandlungen*, ed. by D. Hilbert, 2 Vols. Leipzig 1911. (Chelsea reprint, New York 1967.)
29. Minkowski, H.: Ueber die Bewegung eines festes Körpers in einer Flüsigkeit. Berl. Ber., 1095–1110 (1888)
30. Minkowski, H.: Kapillarität. In: Encyklopädie der mathematischen Wissenschaften mit Einschluss ihrer Anwendungen V, 558–613 (1906)
31. Minkowski, H.: Wärmestrahlung, David Hilbert Nachlass, Göttingen 707 (1907)
32. Minkowski, H.: Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körpern. Gött. Nach., 53–111 (1908) [28, Vol. 2, 352–404]
33. Minkowski, H.: Raum und Zeit. Phys. Zeit. **10**, 104–111 (1909) [28, Vol. 2, 431–444]
34. Minkowski, H.: Das Relativitätsprinzip. Ann. Phys. **47**, 927–938 (1915)
35. Minkowski, H.: Space and Time (Perrett, W., Jeffery, G.B. (English transl) of Minkowski 1909). In: Lorentz et al (eds.) The Principle of Relativity, pp. 73–91. Dover, New York (1952)
36. Norton, J.: Einstein, Nordström and the early demise of scalar, Lorentz-Covariant theories of gravitation. Arch. Hist. Ex. Sci. **45**, 17–94 (1992)
37. Norton, J.: General covariance and the foundations of general relativity. Rep. Prog. Phys. **56**, 791–858 (1993)
38. Pais, A.: Subtle is the Lord. The Science and the Life of Albert Einstein. Oxford University Press, New York (1982)
39. Planck, M.: Das Prinzip der Relativität und die Grundgleichungen der Mechanik. Ver. Deut. Phys. Ges. **8**, 136–141 (1906)
40. Planck, M.: Zur Dynamik der bewegter Systeme. Berl. Ber. **13**, 542–570 (1907) (Repr. in Ann. Phys. **26**, 1–34 (1908))
41. Poincaré, H.: La valeur de la science. Paris (1905)
42. Poincaré, H.: Sur la dynamique de l'électron. Rendiconti del Circolo matimatico di Palermo **21**, 129–176 (1906)
43. Poincaré, H.: Science et méthode, Paris (1908) (English translation: Science and Method, New York, Dover – n.d.)
44. Pyenson, L.: Hermann Minkowski and Einstein's Special Theory of Relativity. Arch. Hist. Ex. Sci. **17**, 71–95 (1977) Repr. in [47, 80–100]
45. Pyenson, L.: Physics in the shadows of Mathematics: The Göttingen Electron-theory seminar of 1905. Arch. Hist. Ex. Sci. **21**, 55–89 (1979) Repr. in [47, 101–136]
46. Pyenson, L.: Relativity in late Wilhelmian Germany: The appeal to a pre-established harmony between Mathematics and Physics. Arch. Hist. Ex. Sci. **24** 138–155 (1982) Repr. in [47, 137–157]
47. Pyenson, L.: The Young Einstein – The Advent of Relativity. Adam Hilger Ltd, Bristol/Boston (1985)
48. Reich, K.: Die Entwicklung des Tensorkalküls. Vom absoluten Differentialkalkül zur Relativitätstheorie. Birkhäuser, Basel/Boston (1994)
49. Reid, C.: Hilbert. Springer, Berlin/New York (1970)
50. Schirrmacher, A.: Experimenting theory: The proofs of Kirchhoff's radiation law before and after Planck. Hist. Stud. Phys. Sci. **33**(2), 299–335 (2003)
51. Schwermer, J.: Räumliche Anschauung und Minima positiv definiter quadratischen Formen. Jahresb. DMV **93**, 49–105 (1991)

52. Seelig, C.: Albert Einstein. Europa Verlag, Zürich (1954)
53. Sommerfeld, A.: Zur Relativitätstheorie. I. Vierdimensionale Vektoralgebra. Ann. Phys. **32**, 749–776 (1910); II. Vierdimensionale Vektoranalysis. Ann. Phys. **33**, 649–689 (1910)
54. Stachel, J.: Einstein and Michelson: The context of discovery and the context of justification. Astronomische Nachrichten **303**, 47–53 (1982)
55. Staley, R.: On the histories of relativity: The propagation and elaboration of relativity theory in participant histories in Germany, 1905–1911. Isis **89**, 263–299 (1998)
56. Walter, S.: Minkowski, mathematicians and the mathematical theory of relativity. In: Goenner, H. et al. (eds.) The Expanding Worlds of General Relativity, pp. 45–86. Birkhäuser, Boston (1999)
57. Walter, S.: The non-Euclidean style of Minkowskian relativity. In: Gray, J.J (ed.) The Symbolic Universe: Geometry and Physics (1990–1930), pp. 91–127. Oxford University Press, New York (1999a)
58. Walter, S.: Breaking in the 4-vectors: Lorentz-covariant gravitation theory, 1905–1910. In: Renn, J., Schemmel, M. (eds.) The genesis of general relativity, Vol. 3, Gravitation in the twilight of classical Physics: Between mechanics, field theory, and astronomy, pp. 193–152. Springer, Berlin (2007)

Chapter 2
Minkowski's Modern World

Scott Walter

Abstract The phenomenal response to Minkowski's 1908 lecture in Cologne has tested the historian's capacity for explanation on rational grounds. What was it about Minkowki's lecture that so schocked the sensibilities of his public? In this essay, Minkowski's spacetime theory is considered as a solution in search of a problem. After physicists rejected his four-dimensional formalism, Minkowski made a point in Cologne of challenging their most cherished beliefs, piling provocation upon provocation in an effort to stir them from their torpor, in pure modernist style.

Keywords Minkowski · Spacetime theory · Relativity theory · History of science · Philosophy of space · Conventionalism

When Hermann Minkowski's first paper on relativity theory [22] appeared in April 1908, it was met with an immediate, largely critical response. The paper purported to extend the reach of the principle of relativity to the electrodynamics of moving media, but one of the founders of relativity theory, the young Albert Einstein, along with his co-author Jakob Laub, found Minkowski's theory to be wanting on physical and formal grounds alike. The lesson in physics delivered by his two former students did not merit a rejoinder, but their summary dismissal of his sophisticated four-dimensional formalism for physics appears to have given him pause.

The necessity of such a formalism for physics was stressed by Minkowski in a lecture entitled "Raum und Zeit," delivered at the annual meeting German Association for Natural Scientists and Physicians in Cologne, on 21 September 1908. Minkowski argued famously in Cologne that certain circumstances required scientists to discard the view of physical space as a Euclidean three-space, in favor of a four-dimensional world with a geometry characterized by the invariance of a certain quadratic form. Delivered in grand style, Minkowski's lecture struck a chord among scientists and philosophers, and upon publication, generated a reaction that was phenomenal in terms of sheer publication numbers and disciplinary breadth.

S. Walter (✉)
University of Nancy and H. Poincaré Archives (CNRS, UMR 7117),
91 av de la Libération, 54001 Nancy, France
e-mail: walter@univ-nancy2.fr

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_2,

Historians have naturally sought to explain this burst of interest in relativity theory. According to one current of thought, Minkowski added nothing of substance to Einstein's theory of relativity, but expressed relativist ideas more forcefully and memorably than Einstein [13, 14]. An alternative explanation claims that Minkowski's explicit appeal to "pre-established harmony" between pure mathematics and physics resonated with Wilhelmine scientists and philosophers, just when Leibnizian ideas were undergoing a revival in philosophical circles [32].

In this paper I want to suggest that much of the excitement generated by Minkowski's Cologne lecture among scientists and philosophers arose from an idea that was scandalous when announced on September 21, 1908, but which was soon assimilated, first by theorists and then by the scientific community at large: Euclidean geometry was no longer adequate to the task of describing physical reality, and had to be replaced by the geometry of a four-dimensional space Minkowski called the "world." Such an affirmation engaged implicitly with the Riemann-Helmholtz-Lie-Poincaré problem of space, and flatly contradicted Poincaré's conventionalist philosophy, whereby the geometry assigned to physical space is a matter of choice, not necessity.

Section 2.1 sketches the background to physical geometry at the time of Minkowski's first lecture on relativity in 1907, and in Section 2.2, the emergence and evolution of the concept of the "world" in Minkowski's writings is discussed, along with a reconstruction of the related discovery of "worldlines." In Section 2.3, the reaction sustained by Minkowski's radical worldview on the part of a few of his most capable readers in physics is reviewed.

2.1 The Geometry of Physical Space Circa 1907

For the few who had followed advances in the electrodynamics of moving bodies up to 1907, including the papers on this topic by Lorentz, Poincaré and Einstein, in Dutch, English, French, and German, the sources of confusion were many and varied about what was physically significant in these theories and what was not. For example, Lorentz employed a coordinate transformation that was meant to be composed with a Galilean transformation, where Poincaré and Einstein folded the two steps into a single transformation, which Poincaré called the "Lorentz" transformation. Poincaré referred to primed and unprimed Lorentz transformations corresponding to motion and relative rest, but within a single frame of reference [42], obviating recourse to the synchronization of clocks, a topic central to Einstein's presentation of relativistic kinematics.[1]

On other points, there was obvious agreement between the first three relativists. For example, all agreed that bodies in motion undergo a certain contraction in the direction of motion; this was the well-known Lorentz-FitzGerald contraction. For

[1] This account draws on standard histories of the special theory of relativity [4, 21].

Poincaré and Einstein, the law of velocity composition was such that the speed of light in vacuum was a maximal velocity, in contradiction with classical mechanics, which features no such speed limit.

The law of velocity composition was a sticking point for physicists, according to one observer, who described it as a "strange result" of Einstein's theory [39]. Einstein derived his law directly from the Lorentz transformation, and expressed it as follows:

$$U = \frac{\sqrt{(v^2 + w^2 + 2vw\cos\alpha) - \left(\frac{vw\sin\alpha}{V}\right)^2}}{1 + \frac{vw\cos\alpha}{V^2}}, \qquad (2.1)$$

where v and w are the velocities to compose, α the angle formed by the velocities, and V the velocity of light [5]. Einstein noted that the parallelogram law of classical kinematics was now valid only in first-order approximation. The focus on a limit relation with the Newtonian (Euclidean) case of the addition law was typical of Einstein's reasoning.

Poincaré, on the other hand, was known to be more of a conqueror than a colonizer in science, and this reputation is borne out by his contribution to relativity theory [49]. For example, Poincaré observed that a Lorentz transformation is a rotation in a four-dimensional vector space with coordinates x, y, z, and $t\sqrt{-1}$ [29]. He used this knowledge to form quadruplets equivalent to modern four-vectors of radius, velocity, force and force density, for application in a Lorentz-invariant law of gravitation. Remarkably, the details of his derivation show that he did not approach his quadruplets as so many directed four-vectors, but as simple Lorentz-invariant quantities. In a word, when Poincaré introduced his four-dimensional vector space, he was not thinking primarily in terms of modern four-vectors [48].

Another important feature of relativity theory noted by Poincaré was the significance of Lorentz's electron theory for classical length measurement. Poincaré asked rhetorically how we go about measuring, and answered as follows [29, p. 132]:

> The first response will be: we transport objects considered to be invariable solids, one on top of the other. But that is no longer true in the current theory if we admit the Lorentzian contraction. In this theory, two lengths are equal, by definition, if they are spanned by light in equal times.

Unlike the standard (Helmholtzian) definition of length congruence based on the free mobility of solids, length congruence in Lorentz's theory depends on the light standard. What Poincaré pointed out, albeit obliquely, was a conflict between the traditional notion of rigidity and the principle of relativity. There are, in fact, no rigid rods in Poincaré's theory of relativity, in stark contrast with Einstein's theory.

According to the doctrine of physical space Poincaré developed in the 1890s, the fact that geometry is an abstract science precludes any knowledge of the geometry of physical space, since the identification of geometric objects (points, lines, planes) with physical processes (lightrays, axes of rotation of regular solids) is arbitrary. His view was essentially equivalent to that of Helmholtz, who recognized the possibility of constructing a non-Newtonian physics based on hyperbolic geometry. But in contrast to Helmholtz, Poincaré insisted on the impossibility of an empirical foundation of the geometry of space, and predicted that Euclidean geometry would forever remain the most convenient geometry [50].

Poincaré recognized as early as 1898 that time and simultaneity were not absolutely given by phenomena, and noted several practical methods of clock synchronization, including clock transport and the exchange of telegraphic signals [27]. Most notably, Poincaré had the genial idea in 1900 of defining operationally Lorentz's "Ortszeit" or local time, as the first-order result of clock synchronization via light signals for two observers relatively at rest, in common motion with respect to the ether, and assuming light isotropy but ignoring the common motion.[2] For Poincaré, his operational definition of local time imbued it with physical meaning. Nonetheless, in his view, local time remained distinct from the "true" time kept by clocks at rest with respect to the ether. And although the local time definition mixed quantities of length and time (using modern notation):

$$t' = t - \frac{\boldsymbol{v}\boldsymbol{x}}{c^2},$$

where $\boldsymbol{v}$ is the frame velocity, $\boldsymbol{x}$ is the spatial separation of the two clocks, and t is the general (ether) time, Poincaré did not perceive any threat here to his conventionalist doctrine of physical space.

In the first years of the twentieth century, many theorists in electrodynamics were familiar with Poincaré's doctrine of physical space and operational definition of local time. That Einstein should employ a synchronization procedure identical to Poincaré's in his first relativity paper is quite natural. In a letter to his friend Habicht, Einstein wrote that his theory involved a "modification of the theory of space and time" [7, Doc. 26]. From kinematic assumptions and light-speed invariance, after much calculation both fastidious and subtle [20], Einstein managed to derive the Lorentz transformation, thereby setting his new physics of inertial frames on sure logical ground.

Einstein's remark to Habicht suggested that his relativity paper would modify the theory of time and space, and the kinematic section of his paper certainly lives up to this billing. The young Einstein had no fear of challenging received wisdom, noisily dismissing from physics the concept of "light-ether," or "absolutely stationary space," the introduction of this concept being "superfluous" [5]. For the rest, Einstein proceeded as if the notion of rigid rods could be applied freely in relativity theory.

As for the notion of time, it did not escape Einstein's attention that it was a path-dependent quantity in his theory. A clock transported with constant speed v around a closed curve, Einstein predicted, would show a lag of $\frac{1}{2}t(v/c)^2$ seconds with respect to a clock at rest, initially synchronized with the mobile clock. For Einstein, this was

[2] For a derivation, see [3]. Poincaré's discovery is linked to his activities as a member of the Bureau of Longitudes in [9]; transcriptions of related letters and reports may be consulted at the Poincaré Correspondence website (www.univ-nancy2.fr/poincare/chp/).

just another "peculiar consequence" of his kinematic assumptions, offering him no further insight to the theory of space and time [5].[3]

How did scientists respond to the theories of Poincaré and Einstein? Most ignored their discoveries, which seemed at first to concern only electron dynamics. In addition, it appeared that the so-called "Lorentz-Einstein" theory was inconsistent with the latest experimental results.

Late in the year 1907, a young mathematician at MIT, with J.W. Gibbs the co-author of an influential treatise on vector analysis, E.B. Wilson complained [53] that while over the previous forty months physicists had taken "long strides" along the path of electron theory, the mathematical theory of electricity had advanced "comparatively little" since the landmark work of Lorentz. Like many scientists of the time, Wilson understood the measurements of electron deflection by the Bonn experimental physicist Walter Kaufmann to have ruled out Lorentz's contractile electron. For Wilson, who was a critic of Poincaré's conventionalist approach to physics, Kaufmann's results held a somewhat deeper meaning both for physics, and for our knowledge of the universe in general, because as Poincaré's work had shown, without the Lorentz electron "the principle of relativity cannot subsist."[4] Although Poincaré had expressed disdain for a tangible ether, and a certain attachment to the relativity of space [30], Wilson was keen to be done with the principle of relativity. "It is certainly more satisfactory philosophically and scientifically," Wilson wrote, "to be left with the hope that some day we may be able to distinguish absolute motion than to feel that we shall in nowise be able to do so."

Wilson's attachment to absolute space was shared by Lorentz, and by most physicists circa 1907, very few of whom had heard of Einstein's theory of relativity, and even fewer of whom who had engaged with it. In all of 1906, nine individuals published on relativity, and by the end of 1907, the yearly total came to 23. In September 1907, when Einstein, then employed as a patent examiner in Bern, was asked by Johannes Stark to write a review article on relativity, Einstein averred acquaintance with only five papers (by four authors) on the topic, excluding his own work.[5]

A month later, Minkowski wrote to Einstein from Göttingen to request an offprint of his first relativity paper [7, Doc. 62]. According to the letter, Minkowski's immediate objective was to prepare a seminar on the partial differential equations of mathematical physics at the University of Göttingen, co-directed with his colleague and best friend, the mathematician David Hilbert [2]. But less than a month later, on November 5 1907, Minkowski delivered a report [24] on Poincaré's theory of gravitation [29] and Planck's recent paper on relativistic dynamics [26] to the Göttingen mathematical society, in which he described his own four-dimensional program for physics, based in part on Einstein's theory.

---

[3] Einstein tacitly assumed that the mobile clock rate depends only on the first derivative of its position vector with respect to a clock at rest [37, p. 68].

[4] Wilson's understanding of the consequence of Kaufmann's results for the principle of relativity was shared by all relativists, including Einstein, at least until Laue [18] recast the dynamics of the contractile electron in four-dimensional terms. For a review of Laue's analysis, see [15].

[5] Two of the five papers mentioned by Einstein appeared before his own writings on relativity. Einstein to Johannes Stark, 25 Sept 1907 [7, Doc. 58].

2.2 World-Geometry (1907–1908)

From the very start of his lecture to the Göttingen mathematical society, Minkowski announced his intellectual gambit: to replace Euclidean geometry of space and time with a certain four-dimensional space. "The world in space and time," Minkowski claimed in his opening remarks, "is, in a certain sense, a four-dimensional non-Euclidean manifold."[6] With hindsight, we might imagine the four-dimensional manifold in question to be that of Minkowski spacetime: a Riemannian four-manifold with Minkowski metric.[7] What Minkowski meant by a *non-Euclidean* manifold, however, was something else altogether [33]. The tip of a four-dimensional velocity vector w_1, w_2, w_3, w_4, Minkowski explained [24, 373],

> is always a point on the surface
>
> $$w_1^2 + w_2^2 + w_3^2 + w_4^2 = -1 \tag{2.2}$$
>
> or, if you wish, on
>
> $$t^2 - x^2 - y^2 - z^2 = 1, \tag{2.3}$$
>
> and represents at the same time the four-dimensional vector from the origin to this point, and this also corresponds to null velocity, to rest, a genuine vector of this sort. Non-Euclidean geometry, of which I spoke earlier in an imprecise fashion, now unfolds for these velocity vectors.

While Minkowski did not bother to unfold the geometry of his velocity vectors, his Göttingen audience would have recognized in (2.2) the equation of a pseudo-hypersphere of unit imaginary radius, and in (2.3) its real counterpart, the two-sheeted unit hyperboloid. The hypersphere (2.2) and the upper sheet ($t > 0$) of the hyperboloid (2.3) had both been popularized by Helmholtz as models of hyperbolic space [11, Vol. 2].

Minkowski observed that the conjugate diameters of the hyperboloid (2.3) give rise to a geometric image of the Lorentz transformation. Any point on (2.3) can be considered to be at rest, in that it may be taken to lie on a t-diameter. This change of axes corresponds to an orthogonal transformation of both the time and space coordinates which is a Lorentz transformation (putting $c = 1$). In modern terms, the three-dimensional hyperboloid (2.3) embedded in four-dimensional Minkowski space affords an interpretation of the Lorentz transformation. This is one of the geometrical insights that was exploited some time later in the form of a spacetime diagram (Fig. 2.1).

A point of interest here is that Minkowski presented the "world in space and time" as a three-dimensional hyperbolic hypersurface. This world of Minkowski's

[6] The published version of Minkowski's talk [24] differs in several key points from the archival typescript (Handschriftenabteilung, Niedersächsische Staats- und Universitätsbibliothek, Göttingen, Math. Archiv 60:3), and excises the concluding paragraph, as noted by Galison [8].

[7] The space Minkowski referred to in this lecture was a certain submanifold of R^4 (actually, the manifold of orthogonal space coordinates x, y, z and a time coordinate t), formed by pairs of quadruplets (x, y, z, t) for which the quadratic form $x^2 + y^2 + z^2 - c^2t^2$ was invariant under an unspecified real linear transformation.

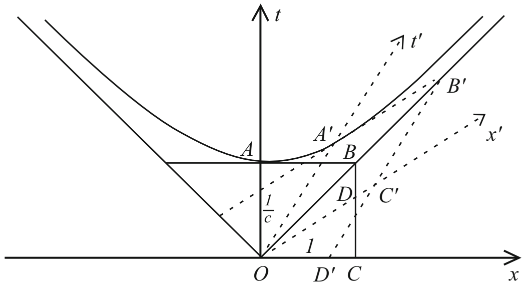


Fig. 2.1 Minkowski's spacetime diagram [23]

was not just another abstract representation of phenomenal space, but physical space itself, even though he felt he had to qualify the affirmation by inserting "in a certain sense" by hand in his typescript (op. cit., note 6).[8]

This new understanding of the structure of relativistic velocity space was a significant step in the direction Minkowski wanted to move, but further progress was blocked by a flawed definition of four-velocity.[9] Applying the method of generalization from three-component vectors to four-component vectors he had applied to find a four-vector potential, four-current density, and four-force density, Minkowski took over the components of the ordinary velocity vector $\mathfrak{w}$ for the spatial part of four-velocity, and added an imaginary fourth component, $i\sqrt{1-\mathfrak{w}^2}$. This gave him four components of four-velocity, w_1, w_2, w_3, w_4:

$$\mathfrak{w}_x, \quad \mathfrak{w}_y, \quad \mathfrak{w}_z, \quad i\sqrt{1-\mathfrak{w}^2}. \tag{2.4}$$

Since the components of Minkowski's quadruplet do not transform like the coordinates of his vector space x_1, x_2, x_3, x_4, they lack what he knew quite well to be an essential property of a four-vector. His error is an interesting one, as it tells us that he did not yet grasp the notion of four-velocity as the four-vector tangent to the worldline of a particle.[10]

Along with a valid four-velocity vector, Minkowski was also missing a four-force vector. With such a spare stock of four-vectors at his disposal, Minkowski's project of expressing relativistic mechanics in four-dimensional terms could not move forward. From a retrospective standpoint, it is rather striking that Minkowski would characterize his new form of the laws of physics as "virtually the greatest triumph

---

[8] Minkowski was not alone in identifying phenomenal space with hyperbolic space, being joined in this stance a few years later by Einstein's correspondent in Zagreb, the mathematician Vladimir Varičak [47].

[9] Minkowski did not employ such four-vector terminology, which was introduced later by Sommerfeld [40].

[10] For a discussion of likely sources of Minkowski's error, see [48]. To see how the definition of four-velocity follows from the definition of a worldline, let the differential parameter $d\tau$ of a worldline be expressed in Minkowskian coordinates by $-d\tau^2 = dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2$. The four-velocity vector w_μ is naturally defined to be the first derivative with respect to this parameter, $w_\mu = dx_\mu/d\tau$ ($\mu = 1, 2, 3, 4$).

ever shown by the application of mathematics" [24, p. 373], even before he had obtained a working set of four-vectors. The lecture contains several major insights, however, and these probably instilled Minkowski with enough confidence to air his nascent program before the members of the Göttingen mathematical society. For example, Minkowski unveiled what he called a "Traktor," a six-component entity equivalent to a modern antisymmetric, second-rank tensor, which Minkowski used to represent the electromagnetic field in a four-dimensional version of Maxwell's field equations.[11]

Five months passed before the mature form of Minkowski's spacetime theory came to light. Minkowski delivered his new theory to the printer on 21 February 1908, and it appeared in the *Göttinger Nachrichten* on 5 April under the title "The Basic Equations for Electromagnetic Processes in Moving Bodies" [48, p. 219, note 84]. Compared to his November lecture, the new paper contains a number of cognitive breakthroughs, including a valid four-velocity vector, the notion of proper time as the parameter of a hyperline in spacetime, the light-hypercone structure of spacetime, and the four-dimensional equations of motion of a material particle.

It is not clear how Minkowski accomplished these breakthroughs. Did he reread Poincaré's memoir, and realize the correct definitions of four-velocity and four-force were readily available? Did he notice that Poincaré's assumption of a lightlike propagation speed of gravitational action between two points in four-dimensional space could be generalized to obtain a lightcone with origin at the source point? Then again, perhaps reading about path-dependent time in the offprint he'd asked for from Einstein put him on the right track. Or maybe Minkowski worked it all out on his own, by studying the embedding of relativistic velocity space (2.3) in four-dimensional vector space, and thereby obtaining further insight into the structure of this hyperspace. I'll return shortly to the latter conjecture.

The result of Minkowski's labors was a 60-page technical memoir packed with new notation, terminology, and calculation rules, featuring a total of six references, and no figures. The four-vectors Minkowski had defined in his earlier lecture now appeared in a new form, along with a single new differential operator named *lor*. As an immediate consequence of this formal extremism, reading Minkowski's paper was a challenging mathematical endeavor.

Terminology changes in "Basic Equations" concerned the "world" itself: all reference to the "world" vanished from "Basic Equations," and along with it, all explicit reference to the velocity space on which the term had been predicated. Minkowski began, as in the November lecture, with a manifold, R^4, and identified a submanifold corresponding to physically-significant points, which he now called "spacetime points" (*Raum-Zeitpunkte*), and "events" (*Ereignisse*). Minkowski now characterized velocity q in terms of the tangent of an imaginary angle $i\psi$,

$$q = -i \tan i\psi, \tag{2.5}$$

[11] A four-dimensional form of Maxwell's potential equations was given in 1906 by a mathematical physicist at the University of Messina, Roberto Marcolongo [48].

where $q < 1$. He could just as well have employed a real angle with a hyperbolic tangent, $q = \tanh\psi$, but did not, perhaps out of a desire to avoid the taint of non-Euclidean geometry, which was likely to offend physicists. From his earlier geometric interpretation of (2.3), Minkowski kept the idea that every rotation of a t-diameter corresponds to a Lorentz transformation, which he now expressed in terms of the angle ψ:

$$x_1' = x_1,\ x_2' = x_2,\ x_3' = x_3\cos i\psi + x_4\sin i\psi,\ x_4' = -x_3\sin i\psi + x_4\cos i\psi. \tag{2.6}$$

In all likelihood, Minkowski was aware of the connection pointed out by Einstein between composition of Lorentz transformations and velocity composition, even though he never mentioned it in print. In fact, Minkowski neither mentioned Einstein's law of velocity addition, nor expressed it mathematically.

While Minkowski suppressed his earlier appeal to the hyperbolic geometry of velocity vectors, he kept the hypersurface (2.3) on which it was based, and provided a new interpretation of its physical significance. This interpretation represents an important clue to understanding how Minkowski discovered the worldline structure of spacetime.

In the appendix to "Basic Equations" devoted to mechanics, Minkowski rehearsed his geometrical interpretation of (2.3), according to which any point on this surface could be chosen such that the line formed with the origin forms a new time axis, and corresponds to a Lorentz transformation. He defined a "spacetime line" to be the totality of spacetime points corresponding to any particular point of matter for all time t. Obvious as this definition may appear to us, it is missing altogether from his November 5 lecture.

With respect to the new concept of a spacetime line, Minkowski noted that its direction is determined at every spacetime point. Here Minkowski introduced the notion of "proper time" (*Eigenzeit*), τ, expressing the increase of coordinate time dt for a point of matter with respect to $d\tau$:

$$d\tau = \sqrt{dt^2 - dx^2 - dy^2 - dz^2} = dt\sqrt{1-\mathfrak{w}^2} = \frac{dx_4}{w_4}, \tag{2.7}$$

where $\mathfrak{w}^2$ is the square of ordinary velocity, $dx_4 = i\,dt$, and $w_4 = i/\sqrt{1-\mathfrak{w}^2}$, which corrects the flawed definition of this fourth component of four-velocity given by Minkowski in his November 5 lecture (2.4).

It is tempting to suppose that Minkowski was led to the discovery of worldlines and proper time by considering the embedding of the hypersurface (2.3) in four-dimensional spacetime, given that he later expressed the norm of a four-velocity vector in the similar form:

$$\left(\frac{dx}{d\tau}\right)^2 + \left(\frac{dy}{d\tau}\right)^2 + \left(\frac{dz}{d\tau}\right)^2 = \left(\frac{dt}{d\tau}\right)^2 - 1, \tag{2.8}$$

and noted that the components of four-velocity can be defined in terms of proper time:

$$\frac{dx}{d\tau}, \qquad \frac{dy}{d\tau}, \qquad \frac{dz}{d\tau}, \qquad i\frac{dt}{d\tau}. \tag{2.9}$$

From these expressions, it appears that the spacetime line has a tangent at every associated spacetime point, and this tangent corresponds to four-velocity. However, this is *not* how Minkowski presented his discovery in the "Basic Equations." A discursive indication supporting this reconstruction is at hand in Minkowski's description of the "direction" of a given spacetime line, determined at every spacetime point. Also, while Minkowski does not actually tie four-velocity to either (2.3) or (2.8) in "Basic Equations," he does so in the Cologne lecture [23, p. 84], employing yet another form of (2.3):

$$c^2\dot{t}^2 - \dot{x}^2 - \dot{y}^2 - \dot{z}^2 = c^2, \tag{2.10}$$

where $\dot{t}$, $\dot{x}$, $\dot{y}$, $\dot{z}$, denote components of four-velocity. The definition of proper time is essential to (2.8), (2.9), and (2.10), but Minkowski remained coy on its origins, describing it only as a "generalization of the concept of local time formed by Lorentz for uniform motion" [22, p. 100]. More than likely, proper time represented much more than this to him. And almost certainly, he viewed (2.3) as the key to spacetime geometry. In a letter to his good friend, former teacher and colleague Adolf Hurwitz, professor of mathematics in Zurich, Minkowski described the "quintessence of my latest studies" to be the "principle of the hyperbolic world" (Minkowski to Hurwitz, 5 May 1908, Niedersächsische Staats- und Universitätsbibliothek, Math. Archiv 78: 212).

Although Minkowski neglected to connect four-velocity to Einstein's law of velocity addition, others did this for him, beginning with Sommerfeld, who expressed parallel velocity addition as the sum of tangents of an imaginary angle [39], followed by Varičak, who recapitulated Sommerfeld's analysis in terms of hyperbolic functions of a real angle [45]. A mathematician at the University of Zagreb, Varičak launched what's been called the "non-Euclidean style" of Minkowskian relativity [47], characterized by an approach to relativity from the standpoint of hyperbolic geometry, based on Einstein's velocity addition (2.1), and Minkowski's Eqs. (2.5) and (2.6). Considered by Sommerfeld as a rival to his own "Euclidean" spacetime formalism, the non-Euclidean style has seen several revivals in various forms over the past century.[12]

What Minkowski found praiseworthy in Einstein's paper was not the law of velocity addition, but the notion of the relativity of simultaneity. Einstein, Minkowski wrote, understood Lorentz-FitzGerald contraction as being "much more a novel view, imposed by phenomena, of the concept of time" than an "artificial hypothesis" [22, p. 55]. But had his former student really understood the relativity of simultaneity? Minkowski noted in a section of his paper entitled "The concept of time" that

[12] For a recent effort, see [34].

the simultaneity of any two events is indeed relative to the motion of the observer (as Einstein had pointed out), just as it is for three simultaneous events, while the simultaneity of four events is absolute, provided the four corresponding spacetime points do not lie on the same spacelike plane [22, p. 69]. He showed, in other words, that Einstein's vertiginous relativity of simultaneity was both grounded and bounded in his four-dimensional view.

Despite its dense and idiosyncratic symbolic notation, Minkowski's "Basic Equations" incited a quick critical response from two of his former students, Einstein and Laub, who had discovered what they thought was an infelicity in Minkowski's definition of ponderomotive force density.[13] They also found that following Minkowski's formalism required too much effort, and set themselves to translating Minkowski's electrodynamics of moving media in terms of ordinary vector analysis [6, Doc. 51]. One imagines that this came as a disappointment to Minkowski. Another of his former students, Gunnar Nordström followed the same path as Einstein and Laub, by showing in his thesis how to derive Minkowski's field equations using ordinary vector analysis [25].

Einstein and Laub's two papers on Minkowski's theory appeared in the July 7 issue of the *Annalen der Physik*, and constitute the only comment on the "Basic Equations" to be published before the September meeting of the German Association of Natural Scientists and Physicians in Cologne. The lecture Minkowski prepared for the mathematical section of this meeting, judging from its overdone rhetoric, may be considered as a reaction to the brutal treatment his spacetime theory had received in the *Annalen*, as Minkowski glorified the discoveries of pure mathematics. Judging from his appeal to Einstein and Lorentz as immediate forebears in the theory of relativity, however, Minkowski aimed also to convince physicists that his spacetime geometry was not entirely useless [46].

In pursuit of the latter goal, Minkowski retrieved the "world" he'd suppressed from the "Basic Equations," and modified its definition. The world as Minkowski now defined it was no longer the hyperbolic space of velocity vectors, but simply the "manifold of all thinkable systems of values x, y, z, t," or what we might call R^4. The geometric objects introduced in "Basic Equations" were renamed accordingly, such that spacetime points segued into worldpoints, spacetime lines into worldlines, and so on. Henceforth, Minkowski's world was no different from the manifold of classical physics, only the geometry had changed.

The redefinition of the "world" is a telling one, as it moved Minkowski's spacetime theory toward the mainstream of scientific thought, and signalled his interest in capturing an audience of physicists. Further evidence of this strategy is provided by the fact that Minkowski retained the two-sheeted hyperboloid (2.3) of his two earlier writings on relativity, but dropped all mention of non-Euclidean geometry. Suppressing the y and z coordinates, Minkowski illustrated (2.3) graphically. Incorporating the asymptotes of (2.3), interpreted as the set of worldpoints capable of

[13] See [6, Doc. 52], and the editorial note "Einstein and Laub on the electrodynamics of moving media" [6, pp. 503–507].

sending light to the origin, or receiving light from the origin, i.e., forward and aft null hypercones, and a second pair of symmetric axes x' and t', Minkowski provided a simple and appealing graphic model of spacetime (see Fig. 2.1).

The transformations leaving (2.3) invariant, combined with arbitrary displacements of the origin, give rise to the inhomogeneous Lorentz group, or what Minkowski referred to simply as G_c. Minkowski naturally affirmed the laws of physics to be invariant with respect to G_c, and interpreted this invariance with reference to his spacetime diagram (Fig. 2.1). According to Minkowski's interpretation, once time is designated t',

> space must then in this connection necessarily be defined by the three-parameter manifold x', y, z, so that physical laws would be expressed in exactly the same way by means of x', y, z, t' as by means of x, y, z, t.

The significance of this remark was twofold. In the first place, for Minkowski the "world" of physical phenomena was no longer characterized by one space (as in Newtonian mechanics), but by "unendlessly many" spaces. In other words, Minkowski underlined the fact that any particle in motion may be considered to be at rest, and can be used to define a time axis and a constant-time hyperplane passing through the origin normal to this axis. Since the choice of particle is arbitrary, the latter hyperplane is only one of an infinity of Euclidean spaces that we may consider in spacetime geometry.[14]

Secondly, the idea that a given definition of time necessarily entails a certain definition of space clearly contradicted Poincaré's doctrine of physical space. To drive his point home, Minkowski famously observed:

> Three-dimensional geometry becomes a chapter of four-dimensional physics. You see why I stated at the outset that space and time should sink into the shadows and only a world in itself subsist.

Where Poincaré had considered three-dimensional geometry to form an inseparable pair with classical physics [50], and had memorably compared physical science to a library, with the theorist in charge of inventory and cataloguing and the experimentalist in charge of acquisitions [28], Minkowski managed to subvert both images in one fell swoop.

Although Minkowski did not mention Poincaré's name in the course of the Cologne lecture, an omission several physicists found odd [46], the French mathematician's doctrine of physical space was an obvious target for him. Poincaré had argued that time and space are not given to us through phenomena, but are the result of conventions. Cognitively endowed by nature with the general notion of a group, humans had adopted Euclidean geometry because the displacement of solid bodies closely approximates the motions of the Euclid group, Poincaré argued [50].[15]

---

[14] Minkowski argued further that Einstein had not recognized this relativity of space, an argument deemed uncharitable, at best, in [46].

[15] Along with Poincaré, but for different reasons, Felix Hausdorff claimed that there was no means of settling the question of the geometry of physical space [10]. Consequently, Minkowski's affir-

Minkowski's anti-conventionalism extended beyond space to include time, or rather, spacetime. This is what he meant, of course, when he said that both space and time should sink into the shadows: any moving particle could be considered to be instantaneously at rest, and any particle at rest could be considered to be in motion with any sublight velocity. The new views of space and time were not mere scenarios, Minkowski insisted, but were imposed on scientists by the circumstances. This was a bold position to adopt in September 1908, as the theory of relativity was then held to be inconsistent with Kaufmann's cathode-ray deflection measurements, as mentioned in (§ 2.1). Minkowski deftly ignored the latter experiments, and took for granted the compatibility of relativity theory with observation.

What were the circumstances that forced a change in conceptions of space and time? This came down to two items in Minkowski's presentation, one formal, the other empirical. The formal consideration was Lorentz-covariance of the differential wave equation of light propagation in empty space. On the empirical side, Minkowski cited the null result of the Michelson-Morley ether-drift experiment. Both of these items had been acknowledged earlier by Lorentz and Poincaré, neither one of whom thought at the time that the traditional concepts of space and time required an overhaul, or even minor repair. Instead, to address the Michelson-Morley null result the latter two theorists appealed to the contraction of moving bodies in the direction of motion with respect to the absolute ether, known as the Lorentz-FitzGerald contraction. The Lorentz-covariance of the wave equation of light was assured by the principle of relativity, upheld by both Lorentz and Poincaré.

Since Minkowski also upheld the principle of relativity, he naturally focused his attention on what distinguished his view from that of Lorentz and Poincaré: the Lorentz-FitzGerald contraction. He suggested that the contraction need not be considered, as he felt it was in Lorentz's view, a "gift from above" (*als Geschenk von oben*). Rather, he wrote, "the Lorentzian hypothesis is fully equivalent to the new view of space and time, whereby it becomes much more comprehensible." Referring to a spacetime diagram (see Fig. 2.2), Minkowski considered the length of an electron with respect to two inertial frames of reference in parallel motion. The details of his demonstration need not be reproduced here,[16] but if we recall Poincaré's proof of electron contraction from Lorentz covariance [29], it is interesting to note Minkowski's choice of "Lorentz electrons" as the bodies of reference, instead of, say, generic solids. Turning the table on Poincaré, Minkowski deduced Lorentz-FitzGerald contraction of electrons from the geometrical relations of spacetime. The structure of spacetime, in other words, was given epistemic priority over the structure of electrons, and consequently over all of physics, in the prevalent micro-physical reduction.

---

mation of the necessity of spacetime for understanding physics implicitly contradicted Hausdorff's view along with that of Poincaré.

[16] For demonstrations of the relativity of simultaneity, time dilation, and length contraction using a spacetime diagram, see [44, p. 98].

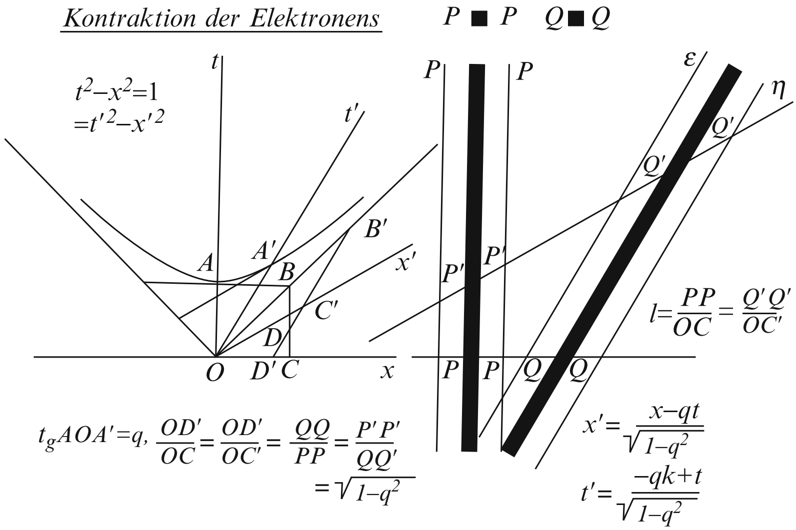


Fig. 2.2 Minkowski's illustration of electron contraction (Courtesy of the Niedersächsische Staats- und Universitätsbibliothek, Göttingen. Math. Arch. 60:2)

2.3 The Reception of "Raum und Zeit"

The text of the lecture "Raum und Zeit" went through at least four drafts, one of which dates from 24 April 1908, before Minkowski submitted it for publication in mid-December 1908. Before the end of 1909, it had appeared in full French and Italian translation and partial English translation, followed by a Russian translation in 1910. Minkowski did not live to see these publications, as he expired on 12 January 1909, following an operation for appendicitis.[17]

The published version of "Raum und Zeit" sparked an explosion of publications in relativity theory, with the number of papers on relativity tripling between 1908 (32 papers) and 1910 (95 papers) [46]. As Hentschel's study showed, the theory of relativity was many things to many people [12], and the same conclusion clearly applies to Minkowski's spacetime theory. For example, a disciplinary analysis of the reception of Minkowski's lecture reveals a overwhelmingly positive response on the part of mathematicians, and a decidedly mixed reaction on the part of physicists [46]. In this section, I suggest that what physicists objected to most in Minkowski's Cologne lecture was the idea that Euclidean space was no longer adequate for understanding physical phenomena.

[17] Minkowski also missed a lecture series by Poincaré, organized by Hilbert for the last week of April in Göttingen. On Poincaré's visit, see the notes in Vol. 2 of the Poincaré Correspondence [31, p. 377], the Hilbert-Poincaré letters (www.univ-nancy2.fr/poincare/chp/), and David Rowe's transcription of two lectures Hilbert delivered for the occasion [35].

Part of the ultimate success of Minkowski's ideas may be attributed to the fact that they were taken up by several members of the Göttingen scientific community, including the titular professors David Hilbert, Felix Klein, Emil Weichert, and former professors Gustav Herglotz and Karl Schwarzschild. In addition, several of Minkowski's former students in Göttingen took up the theory, including Max Born, Max Laue, Theodor Kaluza, Gunnar Nördstrom, Philipp Frank, Emmy Noether, and Ernst Hellinger. As Rowe has remarked [36], the eventual success of Minkowski's theory was a "major triumph" for the Göttingen mathematical community.

Minkowski's Göttingen colleagues had a hand in disseminating Minkowski's work after his death, and in extending its reach in mechanics and electrodynamics, in particular. The individual contributing the most to the success of spacetime geometry among physicists, however, was neither a Göttingen colleague nor a former student, but a friend from Minkowski's schooldays in Königsberg: the Munich theorist Arnold Sommerfeld [1, p. 72]. A former assistant to Klein in Göttingen, and an outsider in theoretical physics, Sommerfeld was initially skeptical of Einstein's relativity, but let himself be won over by Minkowski's theory. Sommerfeld promoted spacetime theory as a technical simplification, which rendered "irrelevant" the "troublesome calculations" of Lorentz and Einstein [46]. Sommerfeld went on to devise a four-dimensional vector algebra and analysis, using symbolic notation and differential operators consonant with those he had imposed on authors as editor of the physics volumes of Klein's vast project, the *Encyclopedia of Mathematical Sciences including Applications* [38]. His streamlined spacetime formalism was taken over and extended by Max Laue, then working in Sommerfeld's institute in Munich, for use in the first German textbook on relativity theory [17]. Laue's textbook was hugely successful, and effectively established the Sommerfeld-Laue formalism as the standard for research in relativity physics.

Sommerfeld insisted upon the simplification afforded to calculation by the adoption of a spacetime approach, and left aside Minkowski's philosophical interpretation of spacetime, with one exception. In the introduction to his 1910 reformulation of Minkowski's matrix calculus, Sommerfeld echoed Minkowski's belief that absolute space should vanish from physics, to be replaced by the "absolute world," by which he meant spacetime geometry, and not R^4 [40, p. 749]. This exchange of absolutes, Euclidean space for Minkowski spacetime, was designed to calm physicists shocked by Minkowski's high-handed dismissal of Euclidean space as the frame for understanding physical phenomena.

Among the shocked physicists was Dantzig's Max Wien, cousin of Willy Wien, the co-editor of *Annalen der Physik*. In a letter to Sommerfeld, Max Wien described his reaction to reading Minkowski's Cologne lecture, which gave him, he wrote, "a slight brain-shiver, now space and time appear conglomerated together in a gray, miserable chaos" [1, p. 71]. Willy Wien was shocked, too, but it wasn't the loss of Euclidean space that bothered him so much as Minkowski's claim that circumstances forced spacetime geometry on physicists. The entire Minkowskian system, Wien said in a 1909 lecture, "evokes the conviction that the facts would have to join it as a fully internal consequence." Wien would have none of this, as he felt that the touchstone of physics was experiment, not abstract mathematical deduction. "For the physicist," Wien concluded his lecture, "Nature alone must make the final decision" [47].

Another physicist, Minkowski's former colleague and director of the Göttingen Institute for Geophysics, Emil Wiechert welcomed Minkowski's spacetime theory, but felt there was no need to dismiss absolute space. Following a remark made by Minkowski in "Raum und Zeit," Wiechert proposed to recover the notion of direction in Euclidean space with what he called "Schreitung" in spacetime, or what amounted to the direction of a four-velocity vector [51]. As for Minkowski's claim that a new intuition of space and time was required, this did not bother Wiechert at all. In a non-technical review of relativity theory [52], Wiechert wrote that the special relativity theory was "brought by Minkowski to a highly mathematically-finished form." He continued:

> It was also Minkowski who, with bold courage, drew the extreme consequences of the theory for a new spacetime-intuition [*Raumzeitanschauung*] and contributed so very much to the theory's renown.

It was precisely Minkowski's spacetime-intuition, or his identification of the extreme consequences of this intuition, that had made the theory of relativity famous in Wiechert's view. For Wiechert, however, all intuitions, including ether, and matter in motion, were but anthropomorphic "images," the reality of which was beyond our ken [51].

A view similar to Wiechert's was expressed by Max Laue in his influential relativity textbook. Laue considered Minkowski spacetime as an "almost indispensable resource" for precise mathematical operations in relativity [17, p. 46]. He expressed reservations, however, about Minkowski's philosophy, in that the geometrical interpretation (or "analogy") of the Lorentz transformation called upon a space of four dimensions:

> [A] geometric analogy can exist only in a four-dimensional manifold. That this is inaccessible to our intuition should not frighten us; it deals only with the symbolic presentation of certain analytical relationships between four variables.

One could avail oneself of the new four-dimensional formalism, Laue assured his readers, even if one was not blessed with Minkowski's spacetime-intuition, and without committing oneself to the existence of Minkowski's four-dimensional world.

By disengaging Minkowski's spacetime ontology from the Sommerfeld-Laue spacetime calculus, Laue cleared the way for the acceptance by physicists of both his calculus, and spacetime geometry in general. A detailed study of the reception of Minkowski's ideas on relativity has yet to be realized, but anecdotal evidence points to a change in attitudes toward Minkowski's spacetime view in the 1950s. For example, in the sixth edition of Laue's textbook, celebrating the fiftieth anniversary of relativity theory, and marking the end of Einstein's life, its author still felt the need to warn physicists away from Minkowski's scandalous claim in Cologne that space and time form a unity. As if in defiance of Laue, this particular view of Minkowski's ("Von Stund' an ...") was soon cited (in the original German) on the title page of a rival textbook on special relativity [43]. In Laue's opinion, however, Minkowski's most famous phrase remained an "exaggeration" [19, p. 60].

Four generations of physicists, dating from the first edition of Laue's textbook, have learned relativity theory in terms of four-vectors and Minkowski maps. Regular application of the Sommerfeld-Laue formalism and spacetime diagrams over the last century has familiarized scientists with spacetime geometry quite thoroughly, making way for numerous variants and extensions. The rules of spacetime diagrams found their way into Feynman diagrams in the 1950s, extending their reach into particle physics, although once again, some physicists objected to their use on the grounds that it was inappropriate to portray trajectories of real particles in this way [16].

2.4 Minkowski's Modern World

Minkowski's carefully-crafted Cologne lecture shocked scientists' sensibilities, in sharp contrast to all previous writings on relativity, including his own. In modernist style, he piled provocation upon provocation: the disciplinary rhetoric, the spate of neologisms, the self-serving (and quite fictional) account of the discovery of spacetime, the anti-conventionalist charge, the discounting of Poincaré, the disingenuous account of Einstein's kinematics, and the geometrical explanation of electron contraction and gravitational attraction, all combine to make "Raum und Zeit" a magnificent example of scientific agitprop.

There was more to the Cologne lecture than mere provocation, of course, as Minkowski took care to place his theory in a distinguished lineage leading from Lorentz, Liénard, Wiechert, Schwarzschild, Einstein and Planck. He claimed not only to have surpassed Lorentz and Einstein, but to have provided a theory of gravitation on an observational par with that of Newton, and to have crafted an electron-based theory of electrodynamics of moving media superior to that of Lorentz. All these claims turned out to be true, adding credibility to the whole, and prompting Arnold Sommerfeld to remark, five years later, that "there is nothing in what Minkowski says that must now be withdrawn." Sommerfeld admitted only one exception: the theory of gravitation, where Einstein's field theory appeared preferable to the action-at-a-distance approaches of Poincaré and Minkowski [41].

The author of "Raum and Zeit" famously characterized his intuitions (*Anschauungen*) of space and time as grounded in experimental physics, and radical in nature. Predictably, his lecture created a scandal for physicists in its day, but unlike most scandals, it did not fade away with the next provocation. Instead, Minkowski focused attention on how mathematics structures our understanding of the physical universe, in a way no other writer had done since Riemann, or has managed to do since, paving the way for acceptance of even more visually-unintuitive theories to come in the early twentieth century, including general relativity and quantum mechanics. Minkowski's provocation of physicists in Cologne, his rejection of existing referents of time, space, and geometry, and his appeal to subjective intuition to describe external reality may certainly be detached from Minkowski geometry, as Laue and others wished, but not if we want to grasp how the concept of spacetime reshaped physics in the early twentieth century.

References

1. Benz, U.: Arnold Sommerfeld: Lehrer und Forscher an der Schwelle zum Atomzeitalter, 1868–1951. Wissenschaftliche Verlagsgesellschaft, Stuttgart (1975)
2. Corry, L.: David Hilbert and the Axiomatization of Physics (1898–1918): From Grundlagen der Geometrie *to* Grundlagen der Physik. Kluwer, Dordrecht (2004)
3. Darrigol, O.: Henri Poincaré's criticism of fin de siècle electrodynamics. Stud. Hist. Philos. M. **26**, 1–44 (1995)
4. Darrigol, O.: Electrodynamics from Ampère to Einstein. Oxford University Press, Oxford (2000)
5. Einstein, A.: Zur Elektrodynamik bewegter Körper. Annalen der Physik **17**, 891–921 (1905) Reed. in [6, Doc. 23]
6. Einstein, A.: The Swiss Years: Writings, 1900–1909. In: Stachel, J., Cassidy, D. C., Renn, J., Schulmann, R. (eds.) The Collected Papers of Albert Einstein, Vol. 2. Princeton University Press, Princeton (1989)
7. Einstein, A.: The Swiss Years: Correspondence, 1902–1914, In: Klein, M.J., Kox, A.J., Schulmann, R. (eds.) The Collected Papers of Albert Einstein, Vol. 5. Princeton University Press, Princeton (1993)
8. Galison, P.: Minkowski's spacetime: From visual thinking to the absolute world. Hist. Stud. Phys. Sci. **10**, 85–121 (1979)
9. Galison, P.: 'Einstein's Clocks and Poincaré's Maps: Empires of Time. Norton, New York (2003)
10. Hausdorff, F.: Das Raumproblem. Annalen der Naturphilosophie **3**, 1–23 (1904)
11. Helmholtz, H.: Vorträge und Reden, 3rd edn. Vieweg, Braunschweig (1884)
12. Hentschel, K.: Interpretationen und Fehlinterpretationen der speziellen und der allgemeinen Relativitätstheorie durch Zeitgenossen Albert Einsteins. Birkhäuser, Basel (1990)
13. Hirosige, T.: Theory of relativity and the ether. Jpn. Stud. Hist. Sci. **7**, 37–53 (1968)
14. Holton, G.: The metaphor of space-time events in science. Eranos Jahrbuch **34**, 33–78 (1965)
15. Janssen, M., Mecklenburg, M.: From classical to relativistic mechanics: Electromagnetic models of the electron. In: Hendricks, V.F., Jørgenson, K.F., Lützen, J., Pedersen, S.A. (eds.) Interactions: Mathematics, Physics and Philosophy, 1860–1930, pp. 65–134. Springer, Dordrecht (2006)
16. Kaiser, D.: Stick-figure realism: Conventions, reification, and the persistence of Feynman diagrams, 1948–1964. Representations **70**, 49–86 (2000)
17. Laue, M.v.: Das Relativitätsprinzip. Vieweg, Braunschweig (1911)
18. Laue, M.v.: Zur Dynamik der Relativitätstheorie. Annalen der Physik **35**, 524–542 (1911)
19. Laue, M.v.: Die Relativitätstheorie: die spezielle Relativitätstheorie 6th edn. Vieweg, Braunschweig (1955)
20. Martínez, A.A.: Kinematics: The Lost Origins of Einstein's Relativity. John's Hopkins University Press, Baltimore (2009)
21. Miller, A.I.: Albert Einstein's Special Theory of Relativity: Emergence (1905) and Early Interpretation. Addison-Wesley, Reading, MA (1981)
22. Minkowski, H.: Die Grundgleichungen für die electromagnetischen Vorgänge in bewegten Körpern. Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, pp. 53–111 (1908)
23. Minkowski, H.: Raum und Zeit. Jahresbericht der deutschen Mathematiker-Vereinigung **18**, 75–88 (1909)
24. Minkowski, H.: Das Relativitätsprinzip. Jahresbericht der deutschen Mathematiker-Vereinigung **24**, 372–382 (1915)
25. Nordström, G.: Die Energiegleichung für das elektromagnetische Feld bewegter Körper. Ph.D. thesis, University of Helsinki (1908)
26. Planck, M.: Zur Dynamik bewegter Systeme. Sitzungsberichte der königliche preußischen Akademie der Wissenschaften, pp. 542–570 (1907)
27. Poincaré, H.: La mesure du temps. Revue de métaphysique et de morale **6**, 1–13 (1898)

28. Poincaré, H.: Les relations entre la physique expérimentale et la physique mathématique. Revue générale des sciences pures et appliquées **11**, 1163–1175 (1900)
29. Poincaré, H.: Sur la dynamique de l'électron. Rendiconti del circolo matematico di Palermo **21**, 129–176, (1906)
30. Poincaré, H.: La relativité de l'espace. Année psychologique **13**, 1–17 (1907)
31. Poincaré, H.: La correspondance entre Henri Poincaré et les physiciens, chimistes et ingénieurs, Walter, S., Bolmont, É., Coret, A. (eds.), La correspondance de Henri Poincaré, Vol. 2. Birkhäuser, Basel (2007)
32. Pyenson, L.: The Young Einstein: The Advent of Relativity. Adam Hilger, Bristol (1985)
33. Reynolds, W.F.: Hyperbolic geometry on a hyperboloid. Am. Math. Month. **100**, 442–455 (1993)
34. Rhodes, J.A., Semon, M.D.: Relativistic velocity space, Wigner rotation, and Thomas precession. Am. J. Phys. **72**, 943–960 (2004)
35. Rowe, D.E.: David Hilbert on Poincaré, Klein, and the world of mathematics. Mathematical Intelligencer **8**, 75–77 (1986)
36. Rowe, D.E.: The Hilbert Problems and the mathematics of a new century. Preprint–Reihe des Fachbereichs Mathematik (1995)
37. Schwartz, H.M.: Introduction to Special Relativity. McGraw-Hill, New York (1968)
38. Sommerfeld, A. (ed.): Physik, vol. 5 of Encyklopädie der mathematischen Wissenschaften mit Einschluss ihrer Anwendungen. Teubner, Leipzig, (1903–1926)
39. Sommerfeld, A.: Über die Zusammensetzung der Geschwindigkeiten in der Relativtheorie. Physikalische Zeitschrift **10**, 826–829 (1909)
40. Sommerfeld, A.: Zur Relativitätstheorie, I: Vierdimensionale Vektoralgebra. Annalen der Physik **32**, 749–776 (1910)
41. Sommerfeld, A.: Anmerkungen zu Minkowski, Raum und Zeit. In: Blumenthal, O. (ed.) Das Relativitätsprinzip; Eine Sammlung von Abhandlungen, pp. 69–73. Teubner, Leipzig (1913)
42. Sternberg, S.: Imagery in scientific thought by Arthur I. Miller. Mathematical Intelligencer **8**, 65–74 (1986)
43. Synge, J.L.: Relativity: The Special Theory. North-Holland, Amsterdam (1956)
44. Torretti, R.: Relativity and Geometry, 2nd edn. Dover, New York (1996)
45. Varičak, V.: Anwendung der Lobatschefskijschen Geometrie in der Relativtheorie. Physikalische Zeitschrift **11**, 93–96 (1910)
46. Walter, S.: Minkowski, mathematicians, and the mathematical theory of relativity. In: Goenner, H., Renn, J., Sauer, T., Ritter, J. (eds.) The expanding worlds of general relativity, vol. 7 of Einstein Studies, pp. 45–86. Birkhäuser, Boston/Basel (1999)
47. Walter, S.: The non-Euclidean style of Minkowskian relativity. In: Gray, J. (ed.) The Symbolic Universe: Geometry and Physics, 1890–1930, pp. 91–127. Oxford University Press, Oxford (1999)
48. Walter, S.: Breaking in the 4-vectors: the four-dimensional movement in gravitation, 1905–1910. In: Renn, J. (ed.) The genesis of general relativity, vol. 3, pp. 193–252. Springer, Berlin (2007)
49. Walter, S.: Poincaré, Jules Henri. In: Koertge, N. (ed.) New Dictionary of Scientific Biography, vol. 6, pp. 121–125. Charles Scribner's Sons, New York (2007)
50. Walter, S.: Hypothesis and convention in Poincaré's defense of Galilei spacetime. In: Heidelberger, M., Schiemann, G. (eds.) The significance of the hypothetical in the natural sciences. De Gruyter, Berlin (2009)
51. Wiechert, E.: Relativitätsprinzip und Äther. Physikalische Zeitschrift **12**, 689–707/737–758 (1911)
52. Wiechert, E.: Die Mechanik im Rahmen der allgemeinen Physik. In: Hinneberg, P., Warburg, E. (eds.) Physik, vol. 3 of Kultur der Gegenwart, pp. 1–78. Teubner, Leipzig (1915)
53. Wilson, E.B. The theory of electricity. Bull. Am. Math. Soc. **14**, 230–237 (1908)

Part II
Implications of Minkowski Spacetime for Theoretical Physics

Chapter 3
Hermann Minkowski and Special Relativity

Graham Hall

Abstract Hermann Minkowski was a mathematical physicist who believed that beauty and elegance were important in the mathematical expression of physical principles. Among his most important works and the work for which his name is, perhaps, best known is his beautiful reformulation of Einstein's special relativity theory. The mathematical ideas and techniques introduced by Minkowski were fundamental in Einstein's construction, a decade later, of the general theory of relativity. This paper explores the mathematical and physical development of "Minkowski Space-Time" and its application to Maxwell's electromagnetic theory and which was presented in two fundamental papers by Minkowski shortly before his death in 1909.

Keywords Minkowski · Space-time · Special relativity

3.1 Introduction

Hermann Minkowski was born on 22 June, 1864, in Alexotas, then in Russia, now the city of Kaunas in Lithuania. His parents were German and, from Alexotas, they moved back to their native land, in fact to Königsberg (now Kaliningrad in Russia), when he was only 8 years old. After attending the Gymnasium in Königsberg he attended the University there from which he was to receive his doctorate in 1885. As a young student, in 1883, he won the Grand Prix of the Academy of Sciences in Paris (sharing the prize with Henry Smith) for his work on number theory. From Königsberg, he moved to a position at the University of Bonn in 1887. In 1896, after a brief return to Königsberg, he was appointed to a position at the Polytechnic in Zurich where he was one of Einstein's teachers. Einstein apparently thought Minkowski an excellent teacher of mathematics. Minkowski's view of Einstein, at the time, was less kind. In 1902 he moved again, this time to a chair at the University of Göttingen, where he remained for the remainder of his relatively short life, dying in 1909 at the age of only 44. In Göttingen, under the influence of David Hilbert, he became interested in mathematical physics.

G. Hall (✉)
Institute of Mathematics, University of Aberdeen, Aberdeen AB24 3UE, Scotland, UK
e-mail: g.hall@abdn.ac.uk

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_3,

In this paper, attention will be focussed on three particular aspects of Minkowski's work in special relativity. These are, first, the four-dimensional reformulation of the geometry and mechanics of Einstein's theory, second, the introduction of the electromagnetic tensor and the consequent rewriting of Maxwell's equations in their most elegant form and third, the construction of the electromagnetic energy-momentum tensor and the four-dimensional description of the conservation laws in Maxwell's theory. In order to realise exactly what Minkowski did in his work on special relativity and, in particular, the dimensionality of "Minkowski Space", it is necessary to discuss some of the general ideas behind not only Einstein's special theory of relativity but also its predecessor, Newtonian theory. In this preamble, some remarks about Hilbert's axiomatisation of Euclidean geometry are relevant. These topics will be dealt with in the next two sections.

3.2 Newtonian Mechanics

In this section, a brief discussion will be given of the general ideas behind the spacetime of classical Newtonian mechanics but described, rather informally, using modern mathematical language. It is important to stress the role played by geometry and, in particular, the impact on theoretical physicists that Minkowski's close friend and colleague, David Hilbert, had with his axiomatisation of Euclidean Geometry. This penetrating development by Hilbert helped to remove much of the obscurity surrounding the concept of "geometry" and paved the way for a clearer understanding of the position of Euclidean geometry and its relationship to the geometry of Lobachevski, Bolyai and others. Such "geometries" were no longer "visual" sciences bedevilled by problems such as what constituted a "point" or a "line" in space, but were controlled by a set of axioms which relegated points, lines and planes to a set of unquestioned and undefined quantities hidden inside various sets. This is, of course, itself open to criticism but, at least, it compartmentalised geometries, clarified the language and gave a firm basis for further discussion which avoided, to some extent, confusion over terminology and hidden implicit assumptions. Thus one can speak of a *Euclidean (geometrical) structure* on a set of the same cardinality as $\mathbb{R}^3$. However, the main point here is whether geometry and physics are related is any serious way. If not, then is the axiomatic formulation of geometry divorced from reality, being a mere idealised arena for the description of natural phonomena, neither conditioned by them nor influencing them in any way? On the other hand, if geometry is concerned with the general ideas of "measurement" and of "rigid" (or, at least, "almost rigid") bodies, then what are the rules that govern their partnership? These two viewpoints were championed, respectively, by Poincaré and Einstein and, of course, the richest such relationship between physics and geometry to date occurs in Einstein's general theory of relativity. This latter theory is outside the scope of this article but it should be mentioned that Minkowski's work in special relativity was a major link between the geometrical ideas of Gauss and Riemann and the application of them to gravitation theory by Einstein. But even in Newtonian theory geometry plays a role since the Galilean law of inertia (Newton's first law) interprets the "straight" lines of Euclidean space geometry as the paths of "free" particles and classical electromagnetic theory reserves such lines for the paths of light rays.

A very brief discussion of Newtonian mechanics can now be given. First one introduces the Universe **U** of all events in space and time together with Newton's *absolute time*, taken as a surjective function $T : \mathbf{U} \to \mathbb{R}$, which records the (absolute) time of each event in **U**. The physical interpretation is that the function T can be simulated by "good clocks" appropriately positioned to give the (absolute) time $T(p)$ of the event $p \in \mathbf{U}$ and which is the same for all observers. The *simultaneity* subsets of **U** are those subsets of **U** of the form $S(t) = T^{-1}(t)$ for $t \in \mathbb{R}$ and each is supposed to be in bijective correspondence with $\mathbb{R}^3$. On each copy of $S(t)$ one assumes a (fixed) Euclidean geometrical structure to have been imposed together with units of measurement of length and angle. Hence the straight lines and planes in each such geometry are represented by the usual linear relationships between the coordinates x, y and z of any Cartesian coordinate system on $S(t)$. An *observer* then consists of a choice of a Cartesian coordinate system on each set $S(t)$. Such a coordinatisation in each $S(t)$ for a given observer then gives rise to a bijective map $g : \mathbf{U} \to \mathbb{R}^4$ for that observer and which assigns to a point $p \in \mathbf{U}$ its x, y and z coordinates and its absolute time $T(p)$. This may be taken as a global chart for **U**, giving **U** the structure of a four-dimensional manifold and which is then *assumed* to be diffeomorphic to the usual manifold $\mathbb{R}^4$ under the diffeomorphism $g : \mathbf{U} \to \mathbb{R}^4$. If another observer gives rise in the same way to a map g' it is assumed, as a result, that the same manifold structure arises on **U** (that is $g \circ g'^{-1}$ and $g' \circ g^{-1}$ are smooth diffeomorphisms $\mathbb{R}^4 \to \mathbb{R}^4$).

Newtonian theory assumes the ability to distinguish between "real" and "inertial" forces and hence it assumes the concept of a *free* particle upon which no real forces act. This leads to the idea of an *inertial* observer as an observer for whom the coordinates x, y and z, chosen in each $S(t)$ as described above, are such that the path of a free particle (a smooth map $\mathbb{R} \to \mathbf{U}$) has a coordinate representation given by $t \to (t, x(t), y(t), z(t))$ where $x(t)$, $y(t)$ and $z(t)$ are *linear* functions of the absolute time t. For a given inertial observer I, a special class of free particles exists whose paths in I's coordinates are such that x, y and z are constant functions. Such particles will be called *fixed points* for I. The collection of fixed points for a particular observer I leads to an identification of the spaces $S(t)$ for I and gives rise to a projection map from **U** to (any) $S(t)$. It then makes sense to speak of the distance *in* I between any two events as the distance between their points of identification in any simultaneity space $S(t)$ (and, of course, it is always sensible to speak of the distance between two points in a simultaneity space, independently of the observer). The projection under this map, with respect to I, of a free particle is then a point or a straight line in (any) $S(t)$. The Galilean law of inertia, in so far as it speaks of straight lines, is then in agreement with the straight lines of the Euclidean geometry originally imposed on the simultaneity spaces. In this sense, *the physics of the law of inertia determines the geometry of the space sections* $S(t)$ (or vice versa). Thus Newtonian theory is described by a *bundle* $(\mathbf{U}, T, \mathbb{R})$ with smooth projection $T : \mathbf{U} \to \mathbb{R}$. This bundle, although diffeomorphic to $\mathbb{R}^4$, has no natural product structure, but each inertial observer I gives rise to obvious *sections* of it through its fixed points. Such sections, and the identifications they give rise to give a product representation of **U** for I and which may be regarded, intuitively, as the *rest (or absolute) space* of I. The smooth relationships between the coordinate systems of inertial observers together with the Newtonian principle of relativity then lead to the Galilean group of transformations. It is remarked that the bundle structure on **U** and, in particular, the fibres from the projection map T, reveal a vestige of three-dimensionality in this (four-dimensional) model. [Of course, if classical electromagnetism is brought into the picture, one must make a claim for the "ether frame" and thus for a preferred observer and associated inertial frame and hence for a preferred product structure on this bundle.]

3.3 Special Relativity

In Special Relativity Theory, one again has the Universe of events **U** but the assumption of the existence of the absolute time function T is now dropped. The Newtonian ideas of real and apparent forces and the concept of a free particle and an inertial frame are retained without change. An inertial observer I is now supposed to possess a "personal" time function in the form of a surjective map $T_I : \mathbf{U} \to \mathbb{R}$ and simultaneity sections $S_I(t) = T_I^{-1}(t)$ for each $t \in \mathbb{R}$. Each set $S_I(t)$ for each I is a copy of $\mathbb{R}^3$ and is given a Euclidean geometrical structure and Cartesian coordinates as in the classical case. The Galilean law of inertia still holds for free particles, using the above coordinates, and fixed points are defined as before. The time function T_I for I is assumed compatible with good clocks *located at the fixed points of* I (and the time function $T_{I'}$ of any other inertial frame I' is assumed compatible with similar clocks *now located at the fixed points of* I'). The fixed points of I enable the simultaneity spaces $S_I(t)$ to be identified as before and thus it again makes sense to speak of the distance, *in* I, between any two events. But now another assumption is required. Let p and q be two events whose distance apart, in an inertial frame I, is d and let $T_I(p) = t_1$ and $T_I(q) = t_2$. Suppose that the events p and q are such that a photon of light can pass through them (and passing first through p in the ordering of T_I). Then this assumption is that there exists a constant c, *the* same *constant for all inertial frames*, such that the distance pq equals $c(t_2 - t_1)$. The constant c is, of course, identified as the *speed of light* and is thus *the same for all inertial observers*. This extra restriction is the main difference between Newtonian theory and special relativity theory and reveals the consistency between the latter theory, the Michelson-Morley (and other similar) experiments and Maxwell's equations. The ether is no longer required. Finally one can impose the principle of relativity in its strongest form to extend the Newtonian one to encompass *all* experiments. Each inertial observer gives, through its space coordinates x, y and z and its time function T_I, a manifold structure to the set **U**. It is assumed, as in the Newtonian case, that these structures are smoothly compatible with each other and hence lead to a unique smooth manifold structure on **U** diffeomorphic to $\mathbb{R}^4$ and to the coordinate transformations between inertial frames which are the Lorentz and Poincaré groups.

These transformations make it clear that the usual Galilean velocity addition laws no longer hold (as indeed they cannot in order that the theory be consistent with the speed of light assumption). In particular, the time transformation is not the trivial one as in Newtonian theory, but a more complicated one which shows quite clearly that the sets $S_I(t)$ and $S_{I'}(t')$ for inertial observers I and I' do not coincide, in general, for any t and t'. Thus there is no observer independent concept of simultaneity and the lack of the bundle structure established in classical theory reveals that the three-dimensional structure still partly apparent in this latter theory is now absent. In this sense, special relativity is four-dimensional. However, the set **U**, regarded as the manifold $\mathbb{R}^4$, has another structure which although commented on by Einstein in [1] was to be explored in more detail by Minkowski.

Suppose one has inertial frames I and I' chosen such that the events in **U** with coordinates $(0, 0, 0, 0)$ in I and $(0, 0, 0, 0)$ in I' are the same event. This may always be arranged by appropriate use of the freedom in the choice of the coordinates and hence of the inertial frames. Call this event O. Now let p be another event in **U** with coordinates $(x,\ y,\ z,\ t)$ in I and coordinates $(x',\ y',\ z',\ t')$ in I'. Consider the quadratic expressions d and d' given for I and I', respectively, by

$$d = x^2 + y^2 + z^2 - c^2t^2 \qquad d' = x'^2 + y'^2 + z'^2 - c^2(t')^2 \tag{3.1}$$

If $d = 0$, a consideration of the first equation above, in I, shows that the events p and O are such that a photon of light can pass through them. But then a consideration of this photon's motion in I' gives $d' = 0$, and conversely. So $d = 0 \Leftrightarrow d' = 0$. This remark is in Einstein's original paper [1] (as a footnote) but he appears not to have considered it any further there. It will be investigated in the next section. Einstein did, however, go on to develop much of what is now regarded as conventional special relativity theory, being the first to give a natural interpretation of the Lorentz transformations and to fully understand the nature of the time transformation law linking the time functions T_I and $T_{I'}$ for two inertial observers I and I'. He also dealt with the velocity addition laws and the Doppler shift and aberration formulae, in addition to the transformation laws applying to Maxwell's equations.

3.4 Minkowski and Special Relativity; Geometry and Kinematics

At this point Minkowski's work can be introduced. His contribution consists largely of two papers [2, 3], the first being a lengthy paper in German and the second a shorter one in English translation an alternative translation of which appears in the present volume. (Regrettably, the former paper does not seem to exist in English translation.) Returning to the quadratic forms introduced in (3.1) and considering the set of all inertial frames with common origin the event O, the quadratic forms d and d' for any two of these frames share their zeros and it is then easily checked that d and d' are proportional. Using the principle of relativity in the form of the indistinguishability of such frames, the proportionality constant must equal unity and so these quadratic forms are equal. Indeed, the fact that the transformations now referred to as the *Lorentz transformations* are essentially those for which the replacement of the primed coordinates in terms of the unprimed ones in the expression d' yields the expression d, was recognised both by Einstein and Minkowski. Thus one can think of **U** in two ways; firstly as the manifold $\mathbb{R}^4$ with a global metric which, in the coordinates x^0, x^1, x^2, x^3, where $x^0 = ct$, $x^1 = x$, $x^2 = y$ and $x^3 = z$, has components $\eta_{ab} = diag(-1, 1, 1, 1)$ or as the vector space $\mathbb{R}^4$ admitting that "generalised" inner product which in the standard basis is represented by the quadratic form

$$\sum_{a,b=0}^{3} \eta_{ab} x^a x^b \tag{3.2}$$

The quadratic form (3.2) is then the *Sylvester canonical form* corresponding to signature $(-1, 1, 1, 1)$. [In fact, Minkowski employed the notation x^4 rather than x^0 and with $x^4 = ict$. This use of "complex time" and the consequent reduction of the real quadratic form above to a complex one with signature $(1, 1, 1, 1)$ did not find general favour and has largely died out. When he used real coordinates, Minkowski inclined towards the signature $(1, -1, -1, -1)$. In this article the signature $(-1, 1, 1, 1)$ will be used. It is also remarked here that, in 1906, Poincaré had also considered a metric of this form [4].] Thus Einstein's theory can now be re-interpreted, through Minkowski, as the manifold $\mathbb{R}^4$ (the above universe **U**) on which is imposed the *Minkowski metric* (labelled η) and whose tangent spaces are the vector space $\mathbb{R}^4$ consisting of "space-time vectors" or *4-vectors*, together with the above *Minkowski inner product*. The group of linear transformations on the vector space $\mathbb{R}^4$ which preserve the form (3.2) is then the pseudo-orthogonal group for that dimension and signature and is the *Lorentz group*. When the frames I and I' are in the usual *standard configuration* (so that their corresponding x, y and z axes are parallel, so that the origin of I' moves along the positive direction of the x axis, as observed in I, with speed v and so that the event of coincidence of the space origins I and I' satisfies $t = 0 = t'$) these transformations are given by

$$x' = \gamma(v)(x - vt), \qquad y' = y, \qquad z' = z, \qquad t' = \gamma(v)\left(t - \frac{v}{c^2}x\right) \tag{3.3}$$

or, in index notation, by

$$x'^1 = \gamma(v)\left(x^1 - \frac{v}{c}x^0\right), \qquad x'^2 = x^2, \qquad x'^3 = x^3, \qquad x'^0 = \gamma(v)\left(x^0 - \frac{v}{c}x^1\right) \tag{3.4}$$

where $\gamma(v) = \left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}}$. Any Lorentz transformation can be written as a composition of transformations like (3.4) together with (generalised) rotations and translations and so, for inertial frames in non-trivial relative motion, transformations like (3.4) are essentially all that is required. Minkowski referred to the set **U** as the *world* and its members as *world points*.

This alternative way of viewing Einstein's special relativity theory led Minkowski to an intuitive description of the difference in dimensionality between it and Newtonian theory. Minkowski pointed out that in Newtonian theory, the Galilean transformations between inertial frames, in an obvious notation and with the preservation of absolute time, took the form $t' = t$, with x' y' and z' then being linear functions of x, y, z and t. Thus the time axis lies arbitrarily with respect to the space sections. This feature is clear from the lack of a natural product structure in the bundle description given in Section 3.2. Such an "angle" between the time axis and the simultaneity spaces, of course, has no significance whatsoever, being simply something in the eye of the person sketching the graph of the bundle and nothing more. It is thus clear that Newtonian theory divides into an absolute time together with infinitely many ways of identifying the simultaneity spaces (and which depend on the inertial observer) and which have no geometrical relationship to each other. The time is fixed and the different identifications are changed by the Galilean transformations between the two inertial observers, according to their relative velocity. However, in Einstein's theory, the Minkowski metric applies to space *and* time and it makes sense to speak of the (generalised) "angle" between *any* two 4-vectors through their inner product. In particular, for an inertial frame, the time axis and the space sections are orthogonal with respect to this metric. Thus the time *and* space sections are *each* moved about by the Lorentz transformations and with the above orthogonality preserved. Minkowski also pointed out that if the Lorentz group of transformations preserving (3.2) is denoted by G_c (this group clearly depends on c) then, informally, G_c tends to the corresponding Galilean group as c becomes arbitrarily large. For Minkowski, the principle of relativity became intimately related to the preservation of the "laws of nature" under G_c. Minkowski preferred to call it the *postulate of the absolute world* rather than the principle of relativity.

Minkowski continued to examine the geometry of what will now be called *space-time* rather than space and time (and is now referred to, most appropriately, as *Minkowski space*). He made use of space-time diagrams and introduced, for the inertial observer I, the subset of points of $\mathbf{U}$ satisfying the condition $x^2 + y^2 + z^2 - c^2t^2 = 0$ which he called a *cone* and is now called the *null cone*. He pointed out how these subsets were independent of the observer I (provided they had the same space-time origin) and thus that space-time was divided up intrinsically by these cones. He also introduced the terms *timelike* and *spacelike* for vectors whose components satisfy, respectively, $x^2 + y^2 + z^2 - c^2t^2 < 0$ and $x^2 + y^2 + z^2 - c^2t^2 > 0$ in I and hence in any such inertial frame. For such a cone, he referred to the regions which are now called the *future* and the *past* of O as *after* O and *before* O.

Minkowski is also responsible for the term *world line* as applied to the path of a particle Q in space-time and he also introduced the very useful and natural concept of the *proper time* of Q. He did this, essentially, by constructing the particle's world line in some inertial frame I as the map $t \to \mathbf{s}(t) \equiv (ct, x(t), y(t), z(t))$ and then defining the proper time (up to an arbitrary additive constant, that is, the choice of the zero of proper time) by the symbol τ and given by $\tau = \frac{1}{c}\int |\mathbf{s}'|dt$, where t is the time coordinate in I, a dash means d/dt and $||$ means the modulus with respect to the Minkowski metric. It follows that $d\tau/dt = (1 - \frac{u^2}{c^2})^{\frac{1}{2}}$ (or $dt/d\tau = \gamma(u)$) where $u^2 = u_x^2 + u_y^2 + u_z^2$ and with $u_x = dx/dt$, $u_y = dy/dt$

and $u_z = dz/dt$ the ordinary (3-) velocity components of Q. Thus τ is the generalised "length" of this path. He then used τ as a parameter along such world lines to give a path $\tau \to \mathbf{s}(\tau) \equiv (ct(\tau), x(\tau), y(\tau), z(\tau))$. This has the advantage that each inertial observer can use the same parameter for Q's path rather than each using his own coordinate time t and then suffering the complications of comparing these descriptions of Q using the time transformation law. He then performed the derivative $d/d\tau$ (denoted by an overdot) to introduce the tangent 4-vector $\dot{\mathbf{s}}$ to this curve (what is now called the *4-velocity* U of Q) and next the 4-vector $\ddot{\mathbf{s}}$, called the *4-acceleration* A of Q. (Minkowski simply used the terms *velocity* and *acceleration* for $\dot{\mathbf{s}}$ and $\ddot{\mathbf{s}}$.) The 4-velocity $\dot{\mathbf{s}}$ of Q has components

$$U = \dot{\mathbf{s}} = \gamma(u)(c,\, u_x,\, u_y,\, u_z) \tag{3.5}$$

Minkowski then noticed that the components of this 4-velocity transformed, under a Lorentz transformation from I to I', just as the coordinates did in (3.4), and hence to the corresponding 4-velocity in I'. (In modern language this would mean it transformed as a *contravariant vector*). This is in contrast to the more complicated transformation laws for the ordinary Newtonian (or 3-) velocity $\mathbf{u} = (u_x,\, u_y,\, u_z)$ under Lorentz transformations and which arise from the (special relativistic) addition laws for this velocity. The same is then true for the 4-acceleration. Minkowski could then introduce his inner product on such contravariant 4-vectors. Thus if 4-vectors V and W have components in I given by $V = (V^0, V^1, V^2, V^3)$ and $W = (W^0, W^1, W^2, W^3)$, their inner product, denoted by $< V, W >$, is defined by

$$< V, W >= \sum_{o}^{3} V^a W^b \eta_{ab} = -V^0W^0 + V^1W^1 + V^2W^2 + V^3W^3 \tag{3.6}$$

this definition being independent of the frame I. The use of the "length" parameter τ meant that $< U, U >$ $(= -c^2)$ is constant along Q's world line and so, by differentiating this latter expression with respect to τ, he thus deduced that the 4-velocity and 4-acceleration were *orthogonal* with respect to his inner product, $< U, A >= 0$, (from which he concluded that the 4-acceleration A is spacelike). Next he introduced the *4-momentum* P associated with Q and defined by $P = m_o U$, where m_0 is the *rest mass* of Q, so that, from (3.5), $P = (mc, mu_x, mu_y, mu_z)$, where $m(= \gamma(u)m_0)$ is the *inertial* mass of Q. Again, P is easily seen to be a contravariant vector. Finally he brought in the *4-force* X defined, rather naturally, by $X = dP/d\tau$ and again X is a contravariant vector. Minkowski's introduction of the 4-force allowed him to give a rather neat expression of the equations of motion of a particle in special relativity. In fact, the work of Planck [5] had shown that one should regard Newton's second law, relativistically, as $\mathbf{f} = d\mathbf{p}/dt$ where $\mathbf{f}$ is the 3 dimensional (3-) force and $\mathbf{p} = m\mathbf{u}$ the corresponding (3-) momentum. *Assuming that the rest mass, m_0, of Q is constant*, $X = m_0 A$ and so, since $< U, A >= 0$, $< X, U >= 0$. It then follows from Planck's result and the above definitions that the components of X are given by $X = \gamma(u)(\frac{\mathbf{f}\cdot\mathbf{u}}{c}, f_1, f_2, f_3)$ where f_1, f_2 and f_3

are the components of $\mathbf{f}$ and that $\mathbf{f} \cdot \mathbf{u} = d/dt(c^2 m)$. Since, $\mathbf{f} \cdot \mathbf{u}$ is the work done by the force $\mathbf{f}$ on Q and thus equals the rate of change of Q's energy, this suggests that one should regard the zero component mc of the 4-momentum P as (c^{-1} times) the total energy mc^2 of the particle. Thus the conservation of inertial mass and (3-) momentum usually used to study collision and scattering problems in relativistic mechanics can be simply stated as the conservation of 4-momentum.

3.5 Minkowski and Special Relativity; Electromagnetic Theory 1

Another very important contribution of Minkowski came in [2] and involved his construction of the electromagnetic field tensor and the consequent (and beautiful) rewriting of Maxwell's equations for the electromagnetic field. The history of the development of electromagnetic theory in the period before Maxwell (for a brief review see [6]) can be traced, firstly, through the early experiments with amber and lodestone culminating in the work of Gilbert at the end of the fifteenth century; secondly, through the seventeenth and eighteenth centuries and the experiments of Cabeo, Boyle, Hauksbee, Gray, du Fay, Franklyn, Volta and Coulomb and thirdly, through the exciting and crucial few years around 1820 when the work of Oersted, Biot, Savart, Ampère and Faraday essentially laid down the final experimental basis for Maxwell's work. Maxwell, building on the works of these people and, in addition, of Kelvin, Weber, Green and Stokes, was able to write down the equations which bear his name and which were first published in 1865 [7] and further clarified in his famous two volume treatise [8]. In modern notation (and which differs from that of Minkowski in only unimportant detail) and considering the "vacuum" situation, if $\mathbf{E}$ is the electric field, $\mathbf{B}$ the magnetic field, $\mathbf{j}$ the current density vector field, $\boldsymbol{\rho}$ the charge density, ε_o and μ_o the permittivity and permeability of free space, respectively (and $c^2 = (\varepsilon_o \mu_o)^{-1}$), and ∇ the usual differential operator, then

$$\nabla \cdot \mathbf{E} = \varepsilon_o^{-1} \boldsymbol{\rho}, \tag{3.7}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{3.8}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} + 1/c^2 \, \partial \mathbf{E}/\partial t, \tag{3.9}$$

$$\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t. \tag{3.10}$$

Equation (3.7) arises from Coulomb's law for the force between electric charges, whilst (3.8) is the statement of the absence of free magnetic poles. Equation (3.9) is Maxwell's corrected form of Ampère's circuital law and introduces the Maxwell displacement current. Thus the set of equations above is now consistent with the law of charge conservation given by

$$\partial \boldsymbol{\rho}/\partial t + \nabla \cdot \mathbf{j} = 0 \tag{3.11}$$

Equation (3.10) is Faraday's induction law.

The fields **E**, **B** and **j** had been assumed to behave as ordinary vectors "in space" in the old theory. Minkowski, having recognised the importance of 4-vectors and their associated transformation laws and their superiority over the ordinary (3-)vectors, attempted to construct a similar quantity which would be as well behaved under Lorentz transformations as his space-time 4-vectors were and which would, in some sense, be *equivalent* to the two fields **E** and **B**, together with another such quantity which would play a similar role for the sources **j** and ρ. To achieve this he first introduced a quantity which had six independent components and which was constructed from the six components of the electric and magnetic field. At each point $p \in U$ and in any inertial frame I he constructed a 4×4 skew-symmetric matrix which in present language is usually labelled F. (Minkowski labelled it f.) With some slight changes in notation when compared with Minkowski's (for consistency with the present use of a real time coordinate rather than Minkowski's complex one, and the renumbering of the coordinates x^0, x^1, x^2, x^3, as described earlier) the matrix F is given in terms of the components (E_1, E_2, E_3) and (B_1, B_2, B_3) of **E** and **B**, respectively, by

$$F^{ab} = \begin{pmatrix} 0 & -E_1 & -E_2 & -E_3 \\ E_1 & 0 & -cB_3 & cB_2 \\ E_2 & cB_3 & 0 & -cB_1 \\ E_3 & -cB_2 & cB_1 & 0 \end{pmatrix} \tag{3.12}$$

Secondly he introduced a four component quantity J given by

$$J^a = (c\rho, j_1, j_2, j_3) \tag{3.13}$$

where j_1, j_2 and j_3 are the components of the ordinary current vector **j**. The positions of the indices in (3.12) and (3.13), and elsewhere, will be explained later. A glance at (3.4) together with the well established invariance of charge reveals that J is a 4-vector (the 4-current) under Lorentz transformations since it can be rewritten as $J^a = \gamma(u)\rho_0(c, u_1, u_2, u_3) = \rho_0 U^a$ where ρ_0 is the *proper* charge density, that is, the charge density in the rest frame of the charge sources, u_1, u_2 and u_3 are the components of the 3-velocity of charge (and u the corresponding speed) and U is the 4-velocity of the charges with components U^a. Minkowski then showed, using the transformation relations between the components of **E** and **B** under the transformation (3.4) between I and I' (given in [1], see also [9])

$$\begin{aligned} E'_1 &= E_1, \qquad E'_2 = \gamma(v)(E_2 - vB_3), \qquad E'_3 = \gamma(v)(E_3 + vB_2) \\ B'_1 &= B_1, \qquad B'_2 = \gamma(v)(B_2 + vE_3/c^2), \qquad B'_3 = \gamma(v)(B_3 - vE_2/c^2) \end{aligned} \tag{3.14}$$

that if F is given by (3.12) in I and by (3.12) with primes on each symbol in I' then F transformed in the manner of what would now be variously referred to as a *contravariant two-form*, a *bivector* or a second order contravariant (skew-symmetric) *tensor* and is thus written in components in any inertial frame I in the conventional way (that is, with indices raised) as the *skew-symmetric* tensor F^{ab}. Minkowski actually pointed out that F transformed as the skew-symmetrised outer product of two basis 4-vectors in any inertial frame and thus as a contravariant (skew-symmetric) tensor. He then noticed that Maxwell's equations could be rewritten in any inertial frame in a rather convenient form. In fact, using Einstein's well-known summation convention (which was not available to Minkowski), these equations become

$$\partial F^{ab}/\partial x^a = \frac{1}{c\varepsilon_0} J^b \qquad \partial F_{ab}/\partial x^c + \partial F_{bc}/\partial x^a + \partial F_{ca}/\partial x^b = 0 \tag{3.15}$$

where the component notation J^a indicates that J is a contravariant 4-vector. This first equation is to be read as four equations, one for each choice of $b = 0, ..., 3$, (Minkowski simply wrote the equations out in full) and which, on substituting (3.12) and recalling that $\partial/\partial t = c\partial/\partial x^0$, can be checked to give the four Equations (3.7) and (3.9). The second equation needs a little explanation. In this equation F_{ab} are the corresponding *covariant* components of F obtained by lowering the indices in F^{ab} with the Minkowski metric η in the usual fashion and given by

$$F_{ab}(= F^{cd}\eta_{ca}\eta_{db}) = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & -cB_3 & cB_2 \\ -E_2 & cB_3 & 0 & -cB_1 \\ -E_3 & -cB_2 & cB_1 & 0 \end{pmatrix} \tag{3.16}$$

The second equation in (3.15) is then easily checked to give four equations, one for each choice of the unordered triple (a, b, c) with a, b and c distinct, and which are (3.8) and (3.10). (Since Minkowski used complex time and hence the complex Euclidean metric he had the same tensor F^{ab} in each of the equations in (3.15).) It is remarked that the *identity* $\partial^2 F^{ab}/\partial x^a \partial x^b = 0$ gives $\partial J^a/\partial x^a = 0$ which, from (3.13), is nothing more than the charge conservation law (3.11).

Minkowski had, in fact, discovered two fundamental results here. First, and from the mathematical viewpoint, the (three-dimensional) fields **E** and **B** transform as ordinary 3- vectors in a complicated way as is clear from (3.14) (just as the ordinary 3-velocity components do). But now, just as Minkowski's 4-velocity transforms nicely under Lorentz transformations, the quantity F transforms in a similar elegant fashion. Secondly, and from the physical viewpoint, the electromagnetic tensor F (now called the *Maxwell tensor* and which, whilst appropriately recognising Maxwell, does not do justice to Minkowski) plays the role of the "invariant" (perhaps one should say "covariant") electromagnetic field. In the old theory, the electric and magnetic fields were given separate existences but it is clear that, at some $p \in \mathbf{U}$, one of these fields may vanish in one inertial frame and not in another. Such a problem can not occur for the quantity F because of its tensor nature. The fields **E** and **B** are now seen as (observer dependent non-invariant) pieces of the real (covariant) field F. In this sense, Minkowski unified the electric and magnetic fields of Maxwell's theory into the single quantity F. It is the observer who decides which part is "electric" and which part is "magnetic".

Minkowski then constructed the *dual* tensor $\overset{*}{F}$ corresponding to F and defined in the present notation by the covariant tensor components $\overset{*}{F}_{ab} = \frac{1}{2}\epsilon_{abcd}F^{cd}$, where ϵ is the usual alternating symbol (so that $\epsilon_{0123} = 1$, etc). Thus

$$\overset{*}{F}_{ab} = \begin{pmatrix} 0 & -cB_1 & -cB_2 & -cB_3 \\ cB_1 & 0 & -E_3 & E_2 \\ cB_2 & E_3 & 0 & -E_1 \\ cB_3 & -E_2 & E_1 & 0 \end{pmatrix} \tag{3.17}$$

He then noted that Maxwell's equations (3.8) and (3.10), which are represented by the second equation in (3.15), could be rewritten in an elegant way in terms of the *contravariant* components $\overset{*}{F}{}^{ab}$ of F, where

$$\overset{*}{F}{}^{ab} = \begin{pmatrix} 0 & cB_1 & cB_2 & cB_3 \\ -cB_1 & 0 & -E_3 & E_2 \\ -cB_2 & E_3 & 0 & -E_1 \\ -cB_3 & -E_2 & E_1 & 0 \end{pmatrix} \tag{3.18}$$

as

$$\partial\overset{*}{F}{}^{ab}/\partial x^a = 0 \tag{3.19}$$

The similarity of this equation with the first equation in (3.15) is clearly seen. [In fact it is interesting to note that Minkowski writes down certain (complex) combinations of the electric and magnetic field components which takes him rather close to defining the so-called *complex self-dual* field $\overset{+}{F} = F + i\overset{*}{F}$ corresponding to F and hence to writing Maxwell's equations in the elegant and compact form $\partial\overset{+}{F}{}^{ab}/\partial x^a = (c\varepsilon_0)^{-1}J^b$]. Next he constructed the two so-called *Maxwell invariants*, α and β, where

$$\alpha = F^{ab}F_{ab} \qquad\qquad \beta = \overset{*}{F}{}^{ab}F_{ab} \tag{3.20}$$

noting their invariance under Lorentz transformations and which, in ordinary three-dimensional language, are $2(c^2\mathbf{B}\cdot\mathbf{B} - \mathbf{E}\cdot\mathbf{E})$ and $4c\mathbf{E}\cdot\mathbf{B}$, respectively, where a dot here means the usual three-dimensional inner product. [Thus α and β are, up to a factor of 2, the real and imaginary parts of $\overset{+}{F}{}^{ab}\overset{+}{F}_{ab}$.] It is remarked that the quantity $\overset{*}{F}{}^{ab}\overset{*}{F}_{ab}$, whilst also invariant in this sense, is not independent of α and β since $\overset{*}{F}{}^{ab}\overset{*}{F}_{ab} = -F^{ab}F_{ab}$. [These invariants arise in the following way. Suppose that, at some point $p \in \mathbf{U}$, the bivector F is such that *at least one* of the above invariants is non-zero (the so-called *non-null* case). Then the "mixed index" tensor $F^a{}_b$, when regarded as a matrix, has exactly two distinct real *null eigendirections* at p, so that if l and n are null vectors at p (so that $< l, l > = < n.n > = 0$)

spanning these directions, $F^a{}_b l^b = \lambda l^a$ and $F^a{}_b n^b = -\lambda n^a$ for some real number λ. That the corresponding eigenvalues differ only in sign is a simple consequence of the skew-symmetry of F and the fact that, since l and n have distinct directions, $< l, n > \neq 0$. These two null directions uniquely determine a two-dimensional subspace of the tangent space at p and hence its orthogonal complement. Let x and y be (necessarily spacelike) vectors at p spanning this orthogonal complement. Then it turns out that if the obvious freedom in the scaling of l, n, x and y is used, so that x and y can be chosen as unit spacelike vectors and the (Minkowski) inner product of l and n as unity, and with a chosen orientation for the ordering of these vectors, so that the dual is defined, F and $\overset{*}{F}$ can be written as

$$F_{ab} = \lambda(l_a n_b - n_a l_b) + b(x_a y_b - y_a x_b) \qquad \overset{*}{F}_{ab} = \lambda(x_a y_b - y_a x_b) - b(l_a n_b - n_a l_b) \tag{3.21}$$

for λ as above and b a real number. A simple calculation then shows that $\alpha = -2\lambda^2 + 2b^2$ and $\beta = 4\lambda b$. Thus α and β determine λ and b up to sign in the sense that for some real numbers e and f either $(\lambda, b) = (e, f)$ or $(\lambda, b) = (-e, -f)$. Thus F is determined (up to a sign) by its geometry and the invariants α and β. If the invariants α and β are both zero at p then F has a unique null eigendirection at p spanned, say, by the null vector l, and represents a *null electromagnetic field* (that is a"radiation" field) with propagation (space-time) direction represented by l. In this case F, $\overset{*}{F}$ and l satisfy $F_{ab}l^b = \overset{*}{F}_{ab}l^b = 0$ and in an inertial frame I in which l has components $(1, n_1, n_2, n_3)$, where (n_1, n_2, n_3) is a unit 3-vector giving the (space) direction of the wave in I, one can check that $\mathbf{E} = -c(\mathbf{n} \times \mathbf{B})$ and that $c\mathbf{B} = \mathbf{n} \times \mathbf{E}$. Thus, in three-dimensional language, $\mathbf{E}$, $\mathbf{B}$ and $\mathbf{n}$ are mutually perpendicular (and the consistency with the vanishing of α and β is clear). The bivector F is then determined by l up to a scaling and an angle representing the "polarisation" of the wave.]

One point may be cleared up here. The transformation properties of the quantity F given in (3.14) are taken from the work of Einstein and Lorentz. They were derived essentially by applying the chain rule for differentiation to Maxwell's equations (3.7–3.10) for $\mathbf{E}$, $\mathbf{B}$, etc in an inertial frame I, using (3.4), and then identifying from the resulting Maxwell-like equations in I' the corresponding electromagnetic field components $\mathbf{E}'$, $\mathbf{B}'$, etc. It is clear that $\mathbf{E}'$, $\mathbf{B}'$ etc, then satisfy Maxwell's equations in I', but since the fields $\mathbf{E}$ and $\mathbf{B}$ are defined in terms of their effect on charged particles through the Lorentz force law, one can ask the following question; do the fields $\mathbf{E}'$ and $\mathbf{B}'$ etc correspond to the same *physical situation* in I', that is, do they represent the same *electromagnetic field* as $\mathbf{E}$ and $\mathbf{B}$? The problem is easily resolved by constructing the covariant 4-force vector $G^a = \frac{e}{c}\eta^{ac} F_{ab} U^a$ along the world line of a charged particle Q which has velocity $\mathbf{u}$ and charge e. The 4-vector properties of G follow from those of F and U. That this actually is the 4-force vector can be seen by noting that in any inertial frame,

$$G^a = \frac{\gamma(u)e}{c}(\mathbf{E} \cdot \mathbf{u},\, Y^1,\, Y^2,\, Y^3) \tag{3.22}$$

where $\frac{e}{c}(Y^1, Y^2, Y^3)$ are the components of the (3-)vector $e(\mathbf{E}+\mathbf{u}\times\mathbf{B})$ and which are the components $\mathbf{f}$ of the (3-) Lorentz force on Q and $\mathbf{f}\cdot\mathbf{u} = e\mathbf{E}\cdot\mathbf{u}$ (see the end of Section 3.4). Thus in any other inertial frame I', the 4-force on Q is represented by the "space" components of G'^a obtained from (3.22) by inserting primes on the quantities $\mathbf{E}$, $\mathbf{B}$ and u. But then the (3-) force on Q in I' is represented by the Lorentz force law with primes on the appropriate quantities and so the identification of $\mathbf{E}'$, $\mathbf{B}'$ etc is justified.

3.6 Minkowski and Special Relativity; Electromagnetic Theory 2

In classical electromagnetic theory, there are separate conservation laws for the energy and momentum of a "vacuum" electromagnetic field with electric field $\mathbf{E}$, magnetic field $\mathbf{B}$, charge density ρ and current vector field $\mathbf{j}$. In considering energy, the total rate of doing work by the field on the charges in a certain (space) volume V is (since the magnetic field does no such work) given by $\int_V \mathbf{j}\cdot\mathbf{E}$. Using Maxwell's equation (3.9) to eliminate the term $\mathbf{j}$, the standard vector identity

$$\nabla\cdot(\mathbf{E}\times\mathbf{B}) = \mathbf{B}\cdot(\nabla\times\mathbf{E}) - \mathbf{E}\cdot(\nabla\times\mathbf{B}) \tag{3.23}$$

and Maxwell's equation (3.10), one easily finds that

$$\int_V \mathbf{j}\cdot\mathbf{E} = -\int_V \left[\partial/\partial t\left[\frac{\varepsilon_0}{2}(\mathbf{E}\cdot\mathbf{E} + c^2\mathbf{B}\cdot\mathbf{B})\right] + \varepsilon_0 c^2\nabla\cdot(\mathbf{E}\times\mathbf{B})\right] \tag{3.24}$$

Thus

$$\mathbf{j}\cdot\mathbf{E} = -\partial\sigma/\partial t - \nabla\cdot\mathbf{P} \tag{3.25}$$

where

$$\mathbf{P} = \varepsilon_0 c^2(\mathbf{E}\times\mathbf{B}), \qquad \sigma = \frac{1}{2}[\varepsilon_0(\mathbf{E}\cdot\mathbf{E} + c^2\mathbf{B}\cdot\mathbf{B})] \tag{3.26}$$

The interpretation of this equation is that σ is the *electromagnetic field energy density* and $\mathbf{P}$ is the *energy flow* or *Poynting* vector. Thus the rate of change of σ plus the energy flowing through the surface of V and represented by Poynting's vector (after converting the appropriate volume integral to a surface integral by means of Gauss' theorem) equals the negative of the work done by the field on the charges. For momentum conservation, one uses the Lorentz force law which gives the ordinary three-dimensional force $\mathbf{f}$ on a charge e with 3-velocity $\mathbf{u}$ in the above field as the expression $\mathbf{f} = e(\mathbf{E}+\mathbf{u}\times\mathbf{B})$. Thus the rate of change of the mechanical momentum Φ in a volume V is given by $d\Phi/dt = \int_V(\rho\mathbf{E}+\mathbf{j}\times\mathbf{B})$. Again one uses Maxwell's equations to eliminate the source terms ρ and $\mathbf{j}$. After a similar standard (if slightly longer) calculation, using the identity

$$\frac{1}{2}\nabla(\mathbf{E}\cdot\mathbf{E}) = (\mathbf{E}\cdot\nabla)\mathbf{E} + \mathbf{E}\times(\nabla\times\mathbf{E}) \tag{3.27}$$

together with a similar one for **B** and Maxwell's equation (3.10), one gets

$$d\Phi/dt + \frac{1}{c^2}\partial/\partial t \int_V \mathbf{P} = -\int \nabla\cdot\mathbf{T} \tag{3.28}$$

So

$$\rho\mathbf{E} + \mathbf{j}\times\mathbf{B} + \frac{1}{c^2}\partial\mathbf{P}/\partial t = -\partial T_{\alpha\beta}/\partial x^\beta \tag{3.29}$$

where **T** is the well known (three-dimensional) symmetric *Maxwell stress-energy tensor* given in components by

$$T_{\alpha\beta} = -\varepsilon_0\left[E_\alpha E_\beta + c^2 B_\alpha B_\beta - \frac{1}{2}\delta_{\alpha\beta}(\mathbf{E}\cdot\mathbf{E} + c^2\mathbf{B}\cdot\mathbf{B})\right] \qquad (\alpha,\ \beta = 1,\ 2,\ 3) \tag{3.30}$$

where $\delta_{\alpha\beta}$ is the three-dimensional Kronecker symbol. The interpretation here is that the rate of change of mechanical momentum in V plus the rate of change of the electromagnetic field momentum in V (represented by the second term on the left hand side of (3.28) equals the volume integral of the divergence of the Maxwell stress-energy tensor.

The final contribution of Minkowski to be discussed here can now be introduced and is that concerning the energy and momentum of the electromagnetic field discussed in the previous paragraph [2]. Minkowski defined a space-time tensor M in terms of the electromagnetic field F (whose tensor properties he confirmed) by

$$M_a{}^b = -\varepsilon_0\left[F_{ac}F^{bc} - \frac{1}{4}\delta_a^b(F_{cd}F^{cd})\right] \tag{3.31}$$

By using the definition of $\overset{*}{F}$ in (3.17) and (3.18), (3.31) can be rewritten either in terms of F^*, in which case one finds an expression like (3.31) with F everywhere replaced by F^*, or in terms of F and $\overset{*}{F}$ as

$$M_a{}^b = -\frac{\varepsilon_0}{2}\left(F_{ac}F^{bc} + \overset{*}{F}_{ac}\overset{*}{F}{}^{bc}\right) \tag{3.32}$$

If one writes out the components of M one gets by using the various expressions for F and $\overset{*}{F}$

$$M_a{}^b = \begin{pmatrix} \sigma & c^{-1}P_1 & c^{-1}P_2 & c^{-1}P_3 \\ -c^{-1}P_1 & -T_{11} & -T_{12} & -T_{13} \\ -c^{-1}P_2 & -T_{12} & -T_{22} & -T_{23} \\ -c^{-1}P_3 & -T_{13} & -T_{23} & -T_{33} \end{pmatrix} \tag{3.33}$$

where (P_1, P_2, P_3) are the components of the Poynting vector $\mathbf{P}$. In these expressions the index b labels the columns and a the rows of M. On raising an index with η this gives the *symmetric* tensor

$$M^{ab}(=\eta^{ac}M_c{}^b) = \begin{pmatrix} -\sigma & -c^{-1}P_1 & -c^{-1}P_2 & -c^{-1}P_3 \\ -c^{-1}P_1 & -T_{11} & -T_{12} & -T_{13} \\ -c^{-1}P_2 & -T_{12} & -T_{22} & -T_{23} \\ -c^{-1}P_3 & -T_{13} & -T_{23} & -T_{33} \end{pmatrix} \quad (3.34)$$

A differentiation of (3.32) and use of Maxwell's equations (3.15) and (3.19) and the identity $(\partial F_{ac}/\partial x^b)F^{bc} = \frac{1}{4}\partial(F_{cd}F^{cd})/\partial x^a$ from (3.15) then gives

$$\partial M_a^b/\partial x^b = -\frac{1}{c}F_{ac}J^c \quad (3.35)$$

(Minkowski used the symbol *lor* (for lorentz) for the four-dimensional divergence operator on the left hand side of (3.35).) Now the left hand side of (3.35) can also be evaluated directly from (3.32) using (3.12) and (3.16–3.18). On equating these two expressions one obtains four equations. The first of these, when $a = 0$, is Eq. (3.25) and the second, when $a = 1, 2, 3$, gives (3.29). Thus the tensor M, which is defined naturally in terms of the electromagnetic bivector F, encompasses within it the conservation laws for both energy and momentum and, once again, Minkowski has constructed a natural space-time quantity M, whose "space-time" divergence gives rise to the well known results described above in a most elegant way. Minkowski also noted the tracefree property of M, that is $M_a^a = 0$, which is just the statement that σ equals the three-dimensional trace of the Maxwell stress-energy tensor, easily deducible from (3.30) and (3.26). [There is an algebraic and geometric link between the work of this section and that of the last one. If, at a point $p \in \mathbf{U}$, the electromagnetic field tensor F used in the construction of M is non-null, the tracefree matrix M is diagonalisable over the real numbers with eigenvalues $h, h, -h, -h$, where h is a linear combination of the scalars α and β in (3.20) and where the h- eigenspace and the (-h)-eigenspace are the orthogonal pair of two-dimensional subspaces determined by F at p. Moreover, if F is replaced in (3.31) or (3.32) by $rF + s\overset{*}{F}$, where r and s are real numbers satisfying $r^2 + s^2 = 1$, the tensor M is unchanged. If F is null at p, then M is not diagonalisable but has Segre type $\{(211)\}$ at p with zero eigenvalue and where the non-simple elementary divisor arises from the null direction uniquely determined by F.]

3.7 Conclusions

In this paper an attempt has been made to describe the work of Minkowski as it applies to Einstein's special theory of relativity. In such a short article as this, justice cannot possibly be done to Minkowski or indeed to many other people who contributed so much to Einstein's special theory in this period, especially Poincaré and Lorentz. A more detailed account of the history of such matters can be found in volume 2 of Whittaker's two volume treatise [10] and also in the excellent book by Pais [11]. Minkowski's work in special relativity some of which has been described here attracts the attention in two different ways. For the mathematician, it is beautifully clever without being over elaborate and for the physicist, it is simple, useful and practical. The overall simplicity achieved by Minkowski in writing out Einstein's special relativity theory in four-dimensional language stands as a supreme intellectual achievement. It taught mathematical physicists to look for "reality" in "space-time" and not in "space" or in "time" separately. In doing this, Minkowski solved several vexing problems. As a trivial example consider the situation of an inertial observer I for whom a certain charge is at rest. Then I experiences only an electric (Coulomb) field from this charge and the associated magnetic field is zero. However, for an inertial frame I' in relative motion with respect to I, a non-zero magnetic field is experienced. Another, more significant, concern was raised by Einstein who, in [1], drew attention to the problem which arises in what is, perhaps, Faraday's most famous experiment, the one regarding electromagnetic induction. If a magnet is moved towards a stationary coil of wire, the latter being attached to a galvanometer, a current is recorded by the galvanometer when the magnet is moving, that is, when the magnetic field created by the magnet is changing (see (3.10)). If, on the other hand, this coil is moved towards a stationary magnet, the same effect is observed (for the same relative velocity) but is now attributed to the effect of the (stationary) magnetic field on the moving charges in the conducting material of the coil of wire. Thus whilst the actual *effect* itself, that is the current recorded, depends only on the relative velocity of the magnet and coil, the reasons given for this effect are quite different. In a relativistic theory one would expect to have the same explanation for each. Of course, the above two explanations for this effect differ because they are given by two different inertial observers, in relative motion, and using the quantities **E** and **B**, etc, and these quantities, as has been showed, are personal measurements of the observer rather than being "covariant" electromagnetic quantities. The description should be in terms of the tensor F. Maxwell's equations are given, covariantly, by the first equation of (3.15) and (3.19) and now appear as restrictions on F. The observer can, of course, choose to compile them *in his own frame* by setting $a = 0$ or $a = 1, 2$ and 3, that is, he can choose to give preference to his own coordinate system. Should he do this he will obtain the "personal" Eqs. (3.7–3.10) with emphasis directed to the personal fields **E** and **B**. If two such observers do this then, not surprisingly, their personal descriptions will differ (cf the complicated laws (3.14)). To quote Minkowski [3] *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.* Some criticism has been expressed of the strength of Minkowski's sentiments in this direction. Maybe some philosophers took this statement too seriously. Mathematicians could easily point to the properties of the Minkowski signature (i.e., the Sylvester canonical form) of η or to the orthogonal group for this dimension and signature (that is the Lorentz group) as giving a clear geometrical distinction between timelike and spacelike vectors. There were even some raised eyebrows and worthless remarks about Minkowski's use of the"imaginary time" coordinate $x_4 = ict$. But Minkowski's methods are clear and sound and provide a most elegant approach to special relativity which, from the practical viewpoint, simplify many calculations. But, more importantly, Minkowski's work significantly increases the *understanding* of the subject. The four-dimensionality of special relativity is emphasised and, on comparison with the material of Section 3.2, the difference in "dimensionality" between it and Newtonian theory is highlighted. It was also fundamental for the later development, by Einstein, of general relativity, where more general metrics than Minkowski's are involved, but Minkowski's signature and basic techniques remain. In particular, many of Minkowski's constructions described above can be generalised with useful effect to general relativity theory. One hundred years later, Minkowski's work in [2, 3], with only minor changes of notation, is still the usual teaching material for special relativity. (For an excellent modern account, Rindler's book [12] is recommended.)

Acknowledgements The author wishes to thank David Lonie, Victor Varela and Kashif Abdul for valuable comments on an earlier version of this paper.

References

1. Einstein, A. Ann. der Phys. **17**, 891 (1905)
2. Minkowski, H. Gott Nach **53** (1908)
3. Minkowski, H. English translation of an address given at the 80th assembly of German Natural Scientists and Physicians at Cologne, 1908. (Reprinted in Minkowski, H., Lorentz, H.A., Weyl, H., Minkowski, H. The Principle of Relativity. Dover, New York (1923))
4. Poincaré, H. Rend. Circ. Palermon. XXI, 129 (1906)
5. Planck, M. Verh. d. Deutsch, Phys. Ges. 136 (1906)
6. Hall, G.S. Lecture at the Maxwell Centenary Meeting, University of Aberdeen (2006) Phil. Trans. Roy. Soc. A **366** (2008) 1849–1860. (Theme Issue; James Clerk Maxwell 150 years on)
7. Maxwell, J.C. Phil. Trans. R. Soc. **155** (1865)
8. Maxwell, J.C. A Treatise on Electricity and Magnetism, two volumes, Oxford (1873)
9. Lorentz, H.A. Proc. Acad. Sci. Amsterdam, 6 (1904), (Reprinted in Minkowski, H., Lorentz, H.A., Weyl, H., Minkowski, H. The Principle of Relativity. Dover, New York (1923))
10. Whittaker, E.T. History of the Theories of Aether and Electricity, Vol 2. Nelson, London (1953)
11. Pais, A. Subtle is the Lord. OUP, Oxford (2005)
12. Rindler, W. Introduction to Special Relativity. OUP, Oxford (1982)

Chapter 4
The Rich Structure of Minkowski Space

Domenico Giulini

Abstract Minkowski Space is the simplest four-dimensional Lorentzian Manifold, being globally flat, and hence the simplest model of spacetime – from a General-Relativistic point of view. But this does not mean that it is trivial in any reasonable sense. In fact, it has a very rich structure, parts of which will be spelled out in detail in this contribution.

Keywords Minkowski space · Special relativity · Affine geometry · Symmetries in Physics

4.1 General Introduction

There are many routes to Minkowski space. But the most physical one still seems to me via the law of inertia. And even along these lines alternative approaches exist. Many papers were published in physics and mathematics journals over the last 100 years in which incremental progress was reported as regards the minimal set of hypotheses from which the structure of Minkowski space could be deduced. One could imagine a Hesse-diagram-like picture in which all these contributions (being the nodes) together with their logical dependencies (being the directed links) were depicted. It would look surprisingly complex.

From a General-Relativistic point of view, Minkowski space just models an *empty* spacetime, that is, a spacetime devoid of any material content. It is worth keeping in mind, that this was not Minkowski's view. Close to the beginning of *Raum und Zeit* he stated:[1]

> In order to not leave a yawning void, we wish to imagine that at every place and at every time something perceivable exists.

D. Giulini (✉)
University of Hannover, Institute for Theoretical Physics, Appelstrasse 2, D-30167 Hannover, Germany
e-mail: giulini@itp.uni-hannover.de

[1] German original: "Um nirgends eine gähnende Leere zu lassen, wollen wir uns vorstellen, daß allerorten und zu jeder Zeit etwas Wahrnehmbares vorhanden ist". ([39], p. 2)

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_4,

This already touches upon a critical point. Our modern theoretical view of spacetime is much inspired by the typical hierarchical thinking of mathematics of the late nineteenth and first half of the twentieth century, in which the *set* comes first, and then we add various structures on it. We first think of spacetime as a set and then structure it according to various physical inputs. But what are the elements of this set? Recall how Georg Cantor, in his first article on transfinite set-theory, defined a set:[2]

> By a 'set' we understand any gathering-together M of determined well-distinguished objects *m* of our intuition or of our thinking (which are called the 'elements' of M) into a whole.

Do we think of spacetime points as "determined well-distinguished objects of our intuition or of our thinking"? I think Minkowski felt a need to do so, as his statement quoted above indicates, and also saw the problematic side of it: If we mentally individuate the points (elements) of spacetime, we – as physicists – have no other means to do so than to fill up spacetime with actual matter, hoping that this could be done in such a diluted fashion that this matter will not dynamically affect the processes that we are going to describe. In other words: The whole concept of a rigid background spacetime is, from its very beginning, based on an assumption of – at best – approximate validity. It is important to realise that this does not necessarily refer to General Relativity: Even if the need to incorporate gravity by a variable and matter-dependent spacetime geometry did not exist would the concept of a rigid background spacetime be of approximate nature, *provided we think of spacetime points as individuated by actual physical events*.

It is true that modern set theory regards Cantor's original definition as too naïve, and that for good reasons. It allows too many "gatherings-together" with self-contradictory properties, as exemplified by the infamous *antinomies* of classical set theory. Also, modern set theory deliberately stands back from any characterisation of elements in order to not confuse the axioms themselves with their possible *interpretations*.[3] However, applications to physics require *interpreted* axioms, where it remains true that elements of sets are thought of as definite as in Cantors original definition.

Modern textbooks on Special Relativity have little to say about this, though an increasing unease seems to raise its voice from certain directions in the philosophy-of-science community; see, e.g., [10, 11]. Physicists sometimes tend to address points of spacetime as *potential events*, but that always seemed to me like poetry,[4]

---

[2] German original: "Unter einer 'Menge' verstehen wir jede Zusammenfassung *M* von bestimmten wohlunterschiedenen Objecten *m* unserer Anschauung oder unseres Denkens (welche die 'Elemente' von *M* genannt werden) zu einem Ganzen." ([12], p. 481)

[3] This urge for a clean distinction between the axioms and their possible interpretations is contained in the famous and amusing dictum, attributed to David Hilbert by his student Otto Blumenthal: "One must always be able to say 'tables', 'chairs', and 'beer mugs' instead of 'points', 'lines', and 'planes'. (German original: "Man muß jederzeit an Stelle von 'Punkten', 'Geraden' und 'Ebenen' 'Tische', 'Stühle' und 'Bierseidel' sagen können.")

[4] "And as imagination bodies forth The forms of things unknown, the poet's pen Turns them to shapes, and gives to airy nothing A local habitation and a name." (A Midsummer Night's Dream, Theseus at V,i)

begging the question how a mere potentiality is actually used for individuation. To me the right attitude seems to admit that the operational justification of the notion of spacetime events is only approximately possible, but nevertheless allow it as primitive element of theorising. The only thing to keep in mind is to not take mathematical rigour for ultimate physical validity. The purpose of mathematical rigour is rather to establish the tightest possible bonds between basic assumptions (axioms) and decidable consequences. Only then can we – in principle – learn anything through falsification.

The last remark opens another general issue, which is implicit in much of theoretical research, namely how to balance between attempted rigour in drawing consequences and attempted closeness to reality when formulating once starting platform (at the expense of rigour when drawing consequences). As the mathematical physicists Glance & Wightman once formulated it in a different context (that of superselection rules in Quantum Mechanics):

> The theoretical results currently available fall into two categories: rigorous results on approximate models and approximate results in realistic models. ([48], p. 204)

To me this seems to be the generic situation in theoretical physics. In that respect, Minkowski space is certainly an approximate model, but to a very good approximation indeed: as global model of spacetime if gravity plays no dynamical rôle, and as local model of spacetime in far more general situations. This justifies looking at some of its rich mathematical structures in detail. Some mathematical background material is provided in the Appendices.

4.2 Minkowski Space and Its Partial Automorphisms

4.2.1 Outline of General Strategy

Consider first the general situation where one is given a set S. Without any further structure being specified, the *automorphisms* group of S would be the group of bijections of S, i.e., maps $f : S \to S$ which are injective (into) and surjective (onto). It is called $\mathrm{Perm}(S)$, where 'Perm' stands for "permutations". Now endow S with some structure Δ; for example, it could be an equivalence relation on S, that is, a partition of S into an exhaustive set of mutually disjoint subsets (cf. Appendix 1). The automorphism group of (S, Δ) is then the subgroup of $\mathrm{Perm}(S \mid \Delta) \subseteq \mathrm{Perm}(S)$ that preserves Δ. Note that $\mathrm{Perm}(S \mid \Delta)$ contains only those maps f preserving Δ whose inverse, f^{-1}, also preserve Δ. Now consider another structure, Δ', and form $\mathrm{Perm}(S \mid \Delta')$. One way in which the two structures Δ and Δ' may be compared is to compare their automorphism groups $\mathrm{Perm}(S \mid \Delta)$ and $\mathrm{Perm}(S \mid \Delta')$. Comparing the latter means, in particular, to see whether one is contained in the other. Containedness clearly defines a partial order relation on the set of subgroups of $\mathrm{Perm}(S)$, which we can use to define a partial order on the set of structures. One structure, Δ, is said to be strictly stronger than (or equally strong as) another structure, Δ', in symbols $\Delta \geq \Delta'$, iff[5] the automorphism group of the former is properly contained in (or is equal to) the automorphism group of the latter.[6] In symbols: $\Delta \geq \Delta' \Leftrightarrow \mathrm{Perm}(S \mid \Delta) \subseteq \mathrm{Perm}(S \mid \Delta')$. Note that in this way of speaking a substructure (i.e., one being defined by a subset of conditions, relations, objects, etc.) of a given structure is said to be weaker than the latter. This way of thinking of structures in terms of their automorphism group is adopted from Felix Klein's *Erlanger Programm* [34] in which this strategy is used in an attempt to classify and compare geometries.

This general procedure can be applied to Minkowski space, endowed with its usual structure (see below). We can than ask whether the automorphism group of Minkowski space, which we know is the inhomogeneous Lorentz group ILor, also called the Poincaré group, is already the automorphism group of a proper substructure. If this were the case we would say that the original structure is redundant. It would then be of interest to try and find a minimal set of structures that already imply the Poincaré group. This can be done by trial and error: one starts with some more or less obvious substructure, determine its automorphism group, and compare it to the Poincaré group. Generically it will turn out larger, i.e., to properly contain ILor. The obvious questions to ask then are: how much larger? and: what would be a minimal extra condition that eliminates the difference?

4.2.2 Definition of Minkowski Space and Poincaré Group

These questions have been asked in connection with various substructures of Minkowski space, whose definition is as follows:

Definition 1. Minkowski space of $n \geq 2$ dimensions, denoted by $\mathbb{M}^n$, is a real n-dimensional affine space, whose associated real n-dimensional vector space V is endowed with a non-degenerate symmetric bilinear form $g : V \times V \to \mathbb{R}$ of signature $(1, n-1)$ (i.e., there exists a basis $\{e_0, e_1, \cdots, e_{n-1}\}$ of V such that $g(e_a, e_b) = \mathrm{diag}(1, -1, \cdots, -1)$). $\mathbb{M}^n$ is also endowed with the standard differentiable structure of $\mathbb{R}^n$.

We refer to Appendix 2 for the definition of affine spaces. Note also that the last statement concerning differentiable structures is put in in view of the strange fact that just for the physically most interesting case, $n = 4$, there exist many inequivalent differentiable structures of $\mathbb{R}^4$. Finally we stress that, at this point, we did not endow Minkowski space with an orientation or time orientation.

Definition 2. The **Poincaré group** in $n \geq 2$ dimensions, which is the same as the **inhomogeneous Lorentz group** in $n \geq 2$ dimensions and therefore will be denoted

[5] Throughout we use "iff" as abbreviation for "if and only if".

[6] Strictly speaking, it would be more appropriate to speak of conjugacy classes of subgroups in $\mathrm{Perm}(S)$ here.

by ILor^n, is that subgroup of the general affine group of real n-dimensional affine space, for which the uniquely associated linear maps $f : V \to V$ are elements of the Lorentz group Lor^n, that is, preserve g in the sense that $g\big(f(v), f(w)\big) = g(v, w)$ for all $v, w \in V$.

See Appendix 3 for the definition of affine maps and the general affine group. Again we stress that since we did not endow Minkowski space with any orientation, the Poincaré group as defined here would not respect any such structure.

As explained in Appendix 4, any choice of an affine frame allows us to identify the general affine group in n dimensions with the semi-direct product $\mathbb{R}^n \rtimes \mathsf{GL}(n)$. That identification clearly depends on the choice of the frame. If we restrict the bases to those where $g(e_a, e_b) = \mathrm{diag}(1, -1, \cdots, -1)$, then ILor^n can be identified with $\mathbb{R}^n \rtimes \mathsf{O}(1, n-1)$.

We can further endow Minkowski space with an *orientation* and, independently, a *time orientation*. An orientation of an affine space is equivalent to an orientation of its associated vector space V. A time orientation is also defined trough a time orientation of V, which is explained below. The subgroup of the Poincaré group preserving the overall orientation is denoted by ILor^n_+ (proper Poincaré group), the one preserving time orientation by $\mathsf{ILor}^n_\uparrow$ (orthochronous Poincaré group), and $\mathsf{ILor}^n_{+\uparrow}$ denotes the subgroup preserving both (proper orthochronous Poincaré group).

Upon the choice of a basis we may identify ILor^n_+ with $\mathbb{R}^n \rtimes \mathsf{SO}(1, n-1)$ and $\mathsf{ILor}^n_{+\uparrow}$ with $\mathbb{R}^n \rtimes \mathsf{SO}_0(1, n-1)$, where $\mathsf{SO}_0(1, n-1)$ is the component of the identity of $\mathsf{SO}(1, n-1)$.

Let us add a few more comments about the elementary geometry of Minkowski space. We introduce the following notations:

$$v \cdot w := g(v, w) \qquad \text{and} \qquad \|v\|_g := \sqrt{|g(v, v)|}\,. \tag{4.1}$$

We shall also simply write v^2 for $v \cdot v$. A vector $v \in V$ is called *timelike, lightlike*, or *spacelike* according to v^2 being >0, =0, or <0 respectively. Non-spacelike vectors are also called *causal* and their set, $\bar{\mathcal{C}} \subset V$, is called the *causal-doublecone*. Its interior, $\mathcal{C}$, is called the *chronological-doublecone* and its boundary, $\mathcal{L}$, the *light-doublecone*:

$$\bar{\mathcal{C}} := \{v \in V \mid v^2 \geq 0\}\,, \tag{4.2a}$$

$$\mathcal{C} := \{v \in V \mid v^2 > 0\}\,, \tag{4.2b}$$

$$\mathcal{L} := \{v \in V \mid v^2 = 0\}\,. \tag{4.2c}$$

A linear subspace $V' \subset V$ is called timelike, lightlike, or spacelike according to $g\big|_{V'}$ being indefinite, negative semi-definite but not negative definite, or negative definite respectively. Instead of the usual Cauchy-Schwarz-inequality we have

$$v^2 w^2 \leq (v \cdot w)^2 \quad \text{for span}\{v, w\} \text{ timelike}\,, \tag{4.3a}$$

$$v^2 w^2 = (v \cdot w)^2 \quad \text{for span}\{v, w\} \text{ lightlike}\,, \tag{4.3b}$$

$$v^2 w^2 \geq (v \cdot w)^2 \quad \text{for span}\{v, w\} \text{ spacelike}\,. \tag{4.3c}$$

Given a set $W \subset V$ (not necessarily a subspace[7]), its g-orthogonal complement is the subspace

$$W^{\perp} := \{v \in V \mid v \cdot w = 0,\ \forall w \in W\}\,. \tag{4.4}$$

If $v \in V$ is lightlike then $v \in v^{\perp}$. In fact, $v^{\perp}$ is the unique lightlike hyperplane (cf. Appendix 2) containing v. In this case the hyperplane $v^{\perp}$ is called degenerate because the restriction of g to $v^{\perp}$ is degenerate. On the other hand, if v is timelike/spacelike $v^{\perp}$ is spacelike/timelike and $v \notin v^{\perp}$. Now the hyperplane $v^{\perp}$ is called non-degenerate because the restriction of g to $v^{\perp}$ is non-degenerate.

Given any subset $W \subset V$, we can attach it to a point p in $\mathbb{M}^n$:

$$W_p := p + W := \{p + w \mid w \in W\}\,. \tag{4.5}$$

In particular, the causal-, chronological-, and light-doublecones at $p \in \mathbb{M}^n$ are given by:

$$\bar{\mathcal{C}}_p \;:=\; p + \bar{\mathcal{C}}\,, \tag{4.6a}$$

$$\mathcal{C}_p \;:=\; p + \mathcal{C}\,, \tag{4.6b}$$

$$\mathcal{L}_p \;:=\; p + \mathcal{L}\,. \tag{4.6c}$$

If W is a subspace of V then W_p is an affine subspace of $\mathbb{M}^n$ over W. If W is time-, light-, or spacelike then W_p is also called time-, light-, or spacelike. Of particular interest are the hyperplanes $v_p^{\perp}$ which are timelike, lightlike, or spacelike according to v being spacelike, lightlike, or timelike respectively.

Two points $p, q \in \mathbb{M}^n$ are said to be timelike-, lightlike-, or spacelike separated if the line joining them (equivalently: the vector $p - q$) is timelike, lightlike, or spacelike respectively. Non-spacelike separated points are also called causally separated and the line though them is called a causal line.

It is easy to show that the relation $v \sim w \Leftrightarrow v \cdot w > 0$ defines an equivalence relation (cf. Appendix 1) on the set of timelike vectors. (Only transitivity is non-trivial, i.e., if $u \cdot v > 0$ and $v \cdot w > 0$ then $u \cdot w > 0$. To show this, decompose u and w into their components parallel and perpendicular to v.) Each of the two equivalence classes is a *cone* in V, that is, a subset closed under addition and multiplication with positive numbers. Vectors in the same class are said to have the same time orientation. In the same fashion, the relation $v \sim w \Leftrightarrow v \cdot w \geq 0$ defines an equivalence relation on the set of causal vectors, with both equivalence classes being again cones. The existence of these equivalence relations is expressed by saying that $\mathbb{M}^n$ is *time orientable*. Picking one of the two possible time orientations is then equivalent to specifying a single timelike reference vector, v_*, whose equivalence class of directions may be called the *future*. This being done we can speak of the future (or forward, indicated by a superscript$+$) and past (or backward, indicated by a superscript $-$) cones:

[7] By a "subspace" of a vector space we always understand a sub vector-space.

$$\bar{\mathcal{C}}^{\pm} := \{v \in \bar{\mathcal{C}} \mid v \cdot v_* \gtrless 0\}\,, \tag{4.7a}$$

$$\mathcal{C}^{\pm} := \{v \in \mathcal{C} \mid v \cdot v_* \gtrless 0\}\,, \tag{4.7b}$$

$$\mathcal{L}^{\pm} := \{v \in \mathcal{L} \mid v \cdot v_* \gtrless 0\}\,. \tag{4.7c}$$

Note that $\bar{\mathcal{C}}^{\pm} = \mathcal{C}^{\pm} \cup \mathcal{L}^{\pm}$ and $\mathcal{C}^{\pm} \cap \mathcal{L}^{\pm} = \emptyset$. Usually $\mathcal{L}^{+}$ is called the future and $\mathcal{L}^{-}$ the past lightcone. Mathematically speaking this is an abuse of language since, in contrast to $\bar{\mathcal{C}}^{\pm}$ and $\mathcal{C}^{\pm}$, they are not cones: They are each invariant (as sets) under multiplication with positive real numbers, but adding to vectors in $\mathcal{L}^{\pm}$ will result in a vector in $\mathcal{C}^{\pm}$ unless the vectors were parallel.

As before, these cones can be attached to the points in $\mathbb{M}^n$. We write in a straightforward manner:

$$\bar{\mathcal{C}}^{\pm}_p := p + \bar{\mathcal{C}}^{\pm}\,, \tag{4.8a}$$

$$\mathcal{C}^{\pm}_p := p + \mathcal{C}^{\pm}\,, \tag{4.8b}$$

$$\mathcal{L}^{\pm}_p := p + \mathcal{L}^{\pm}\,. \tag{4.8c}$$

The Cauchy-Schwarz inequalities (4.3) result in various generalised triangle-inequalities. Clearly, for spacelike vectors, one just has the ordinary triangle inequality. But for causal or timelike vectors one has to distinguish the cases according to the relative time orientations. For example, for timelike vectors of equal time orientation, one obtains the reversed triangle inequality:

$$\|v + w\|_g \geq \|v\|_g + \|w\|_g\,, \tag{4.9}$$

with equality iff v and w are parallel. It expresses the geometry behind the "twin paradox".

Sometimes a Minkowski 'distance function' $d : \mathbb{M}^n \times \mathbb{M}^n \to \mathbb{R}$ is introduced through

$$d(p, q) := \|p - q\|_g\,. \tag{4.10}$$

Clearly this is not a distance function in the ordinary sense, since it is neither true that $d(p, q) = 0 \Leftrightarrow p = q$ nor that $d(p, w) + d(w, q) \geq d(p, q)$ for all p, q, w.

4.2.3 *From Metric to Affine Structures*

In this section we consider general isometries of Minkowski space. By this we mean general bijections $F : \mathbb{M}^n \to \mathbb{M}^n$ (no requirement like continuity or even linearity is made) which preserve the Minkowski distance (4.10) as well as the time or spacelike character; hence

$$\big(F(p) - F(q)\big)^2 = (p - q)^2 \qquad \text{for all} \quad p, q \in \mathbb{M}^n\,. \tag{4.11}$$

Poincaré transformations form a special class of such isometries, namely those which are affine. Are there non-affine isometries? One might expect a whole Pandora's box full of wild (discontinuous) ones. But, fortunately, they do not exist: Any map $f : V \to V$ satisfying $(f(v))^2 = v^2$ for all v must be linear. As a warm up, we show

Theorem 1. *Let $f : V \to V$ be a surjection (no further conditions) so that $f(v) \cdot f(w) = v \cdot w$ for all $v, w \in V$, then f is linear.*

Proof. Consider $I := \big(af(u) + bf(v) - f(au + bv)\big) \cdot w$. surjectivity allows to write $w = f(z)$, so that $I = a\,u \cdot z + b\,v \cdot z - (au + bv) \cdot z$, which vanishes for all $z \in V$. Hence $I = 0$ for all $w \in V$, which by non-degeneracy of g implies the linearity of f. □

This shows in particular that any bijection $F : \mathbb{M}^n \to \mathbb{M}^n$ of Minkowski space whose associated map $f : V \to V$, defined by $f(v) := F(o + v) - F(o)$ for some chosen basepoint o, preserves the Minkowski metric must be a Poincaré transformation. As already indicated, this result can be considerably strengthened. But before going into this, we mention a special and important class of linear isometries of (V, g), namely reflections at non-degenerate hyperplanes. The reflection at $v^\perp$ is defined by

$$\rho_v(x) := x - 2\,v\,\frac{x \cdot v}{v^2}\,. \tag{4.12}$$

Their significance is due to the following

Theorem 2 (Cartan, Dieudonné). *Let the dimension of V be n. Any isometry of (V, g) is the composition of at most n reflections.*

Proof. Comprehensive proofs may be found in [31] or [5]. The easier proof for at most $2n - 1$ reflections is as follows: Let ϕ be a linear isometry and $v \in V$ so that $v^2 \neq 0$ (which certainly exists). Let $w = \phi(v)$, then $(v + w)^2 + (v - w)^2 = 4v^2 \neq 0$ so that $w + v$ and $w - v$ cannot simultaneously have zero squares. So let $(v \mp w)^2 \neq 0$ (understood as alternatives), then $\rho_{v \mp w}(v) = \pm w$ and $\rho_{v \mp w}(w) = \pm v$. Hence v is eigenvector with eigenvalue 1 of the linear isometry given by

$$\phi' = \begin{cases} \rho_{v-w} \circ \phi & \text{if } (v - w)^2 \neq 0\,, \\ \rho_v \circ \rho_{v+w} \circ \phi & \text{if } (v - w)^2 = 0\,. \end{cases} \tag{4.13}$$

Consider now the linear isometry $\phi'\big|_{v^\perp}$ on $v^\perp$ with induced bilinear form $g\big|_{v^\perp}$, which is non-degenerated due to $v^2 \neq 0$. We conclude by induction: At each dimension we need at most two reflections to reduce the problem by one dimension. After $n - 1$ steps we have reduced the problem to one dimension, where we need at most one more reflection. Hence we need at most $2(n - 1) + 1 = 2n - 1$ reflections which upon composition with ϕ produce the identity. Here we use that any linear isometry in $v^\perp$ can be canonically extended to $\mathrm{span}\{v\} \oplus v^\perp$ by just letting it act trivially on $\mathrm{span}\{v\}$. □

Note that this proof does not make use of the signature of g. In fact, the theorem is true for any signatures; it only depends on g being symmetric and non degenerate.

4.2.4 From Causal to Affine Structures

As already mentioned, Theorem 1 can be improved upon, in the sense that the hypothesis for the map being an isometry is replaced by the hypothesis that it merely preserve some relation that derives form the metric structure, but is not equivalent to it. In fact, there are various such relations which we first have to introduce.

The family of cones $\{\bar{\mathcal{C}}_q^+ \mid q \in \mathbb{M}^n\}$ defines a partial-order relation (cf. Appendix 1), denoted by $\geq$, on spacetime as follows: $p \geq q$ iff $p \in \bar{\mathcal{C}}_q^+$, i.e., iff $p-q$ is causal and future pointing. Similarly, the family $\{\mathcal{C}_q^+ \mid q \in \mathbb{M}^n\}$ defines a strict partial order, denoted by $>$, as follows: $p > q$ iff $p \in \mathcal{C}_q^+$, i.e., if $p-q$ is timelike and future pointing. There is a third relation, called $\gtrdot$, defined as follows: $p \gtrdot q$ iff $p \in \mathcal{L}_q^+$, i.e., p is on the future lightcone at q. It is not a partial order due to the lack of transitivity, which, in turn, is due to the lack of the lightcone being a cone (in the proper mathematical sense explained above). Replacing the future (+) with the past (−) cones gives the relations $\leq$, $<$, and $\lessdot$.

It is obvious that the action of $\mathsf{ILor}^\uparrow$ (spatial reflections are permitted) on $\mathbb{M}^n$ maps each of the six families of cones (4.8) into itself and therefore leave each of the six relations invariant. For example: Let $p > q$ and $F \in \mathsf{ILor}^\uparrow$, then $(p-q)^2 > 0$ and $p-q$ future pointing, but also $(F(p)-F(q))^2 > 0$ and $F(p)-F(q)$ future pointing, hence $F(p) > F(q)$. Another set of "obvious" transformations of $\mathbb{M}^n$ leaving these relations invariant is given by all dilations:

$$d_{(\lambda,m)} : \mathbb{M}^n \to \mathbb{M}^n\,, \quad p \mapsto d_{(\lambda,m)}(p) := \lambda(p-m)+m\,, \tag{4.14}$$

where $\lambda \in \mathbb{R}_+$ is the constant dilation-factor and $m \in \mathbb{M}^n$ the centre. This follows from $\big(d_{\lambda,m}(p)-d_{\lambda,m}(q)\big)^2 = \lambda^2(p-q)^2$, $\big(d_{\lambda,m}(p)-d_{\lambda,m}(q)\big)\cdot v_* = \lambda(p-q)\cdot v_*$, and the positivity of λ. Since translations are already contained in $\mathsf{ILor}^\uparrow$, the group generated by $\mathsf{ILor}^\uparrow$ and all $d_{\lambda,m}$ is the same as the group generated by $\mathsf{ILor}^\uparrow$ and all $d_{\lambda,m}$ for fixed m.

A seemingly difficult question is this: What are the most general transformations of $\mathbb{M}^n$ that preserve those relations? Here we understand "transformation" synonymously with 'bijective map', so that each transformation f has in inverse f^{-1}. 'Preserving the relation' is taken to mean that f *and* f^{-1} preserve the relation. Then the somewhat surprising answer to the question just posed is that, in three or more spacetime dimensions, there are no other such transformations besides those already listed:

Theorem 3. *Let $\succ$ stand for any of the relations $\geq, >, \gtrdot$ and let F be a bijection of $\mathbb{M}^n$ with $n \geq 3$, such that $p \succ q$ implies $F(p) \succ F(q)$ and $F^{-1}(p) \succ F^{-1}(q)$. Then F is the composition of an Lorentz transformation in $\mathsf{ILor}^\uparrow$ with a dilation.*

Proof. These results were proven by A.D. Alexandrov and independently by E.C. Zeeman. A good review of Alexandrov's results is [1]; Zeeman's paper is [49]. The restriction to $n \geq 3$ is indeed necessary, as for $n = 2$ the following possibility exists: Identify $\mathbb{M}^2$ with $\mathbb{R}^2$ and the bilinear form $g(z, z) = x^2 - y^2$, where $z = (x, y)$. Set $u := x - y$ and $v := x + y$ and define $f : \mathbb{R}^2 \to \mathbb{R}^2$ by $f(u, v) := (h(u), h(v))$, where $h : \mathbb{R} \to \mathbb{R}$ is any smooth function with $h' > 0$. This defines an orientation preserving diffeomorphism of $\mathbb{R}^2$ which transforms the set of lines $u =$ const. and $v =$ const. respectively into each other. Hence it preserves the families of cones (4.8a). Since these transformations need not be affine linear they are not generated by dilations and Lorentz transformations.

These results may appear surprising since without a continuity requirement one might expect all sorts of wild behaviour to allow for more possibilities. However, a little closer inspection reveals a fairly obvious reason for why continuity is implied here. Consider the case in which a transformation F preserves the families $\{\mathcal{C}_q^+ \mid q \in \mathbb{M}^n\}$ and $\{\mathcal{C}_q^- \mid q \in \mathbb{M}^n\}$. The open diamond-shaped sets (usually just called "open diamonds"),

$$U(p, q) := (\mathcal{C}_p^+ \cap \mathcal{C}_q^-) \cup (\mathcal{C}_q^+ \cap \mathcal{C}_p^-)\,, \tag{4.15}$$

are obviously open in the standard topology of $\mathbb{M}^n$ (which is that of $\mathbb{R}^n$). Note that at least one of the intersections in (4.15) is always empty. Conversely, is is also easy to see that each open set of $\mathbb{M}^n$ contains an open diamond. Hence the topology that is defined by taking the $U(p, q)$ as sub-base (the basis being given by their finite intersections) is equivalent to the standard topology of $\mathbb{M}^n$. But, by hypothesis, F and F^{-1} preserves the cones $\mathcal{C}_q^\pm$ and therefore open sets, so that F must, in fact, be a homeomorphism.

There is no such *obvious* continuity input if one makes the strictly weaker requirement that instead of the cones (4.8) one only preserves the doublecones (4.6). Does that allow for more transformations, except for the obvious time reflection? The answer is again in the negative. The following result was shown by Alexandrov (see his review [1]) and later, in a different fashion, by Borchers and Hegerfeld [8]:

Theorem 4. *Let $\sim$ denote any of the relations: $p \sim q$ iff $(p - q)^2 \geq 0$, $p \sim q$ iff $(p - q)^2 > 0$, or $p \sim q$ iff $(p - q)^2 = 0$. Let F be a bijection of $\mathbb{M}^n$ with $n \geq 3$, such that $p \sim q$ implies $F(p) \sim F(q)$ and $F^{-1}(p) \sim F^{-1}(q)$. Then F is the composition of an Lorentz transformation in* ILor *with a dilation.*

All this shows that, up to dilations, Lorentz transformations can be characterised by the causal structure of Minkowski space. Let us focus on a particular sub-case of Theorem 4, which says that any bijection F of $\mathbb{M}^n$ with $n \geq 3$, which satisfies $\|p - q\|_g = 0 \Leftrightarrow \|F(p) - F(q)\|_g = 0$ must be the composition of a dilation and a transformation in ILor. This is sometimes referred to as *Alexandrov's theorem*. It gives a precise answer to the following physical question: To what extent does the principle of the constancy of a finite speed of light *alone* determine the relativity group? The answer is, that it determines it to be a subgroup of the 11-parameter group of Poincaré transformations and constant rescalings, which is as close to the Poincaré group as possibly imaginable.

Alexandrov's Theorem is, to my knowledge, the closest analog in Minkowskian geometry to the famous theorem of Beckman and Quarles [3], which refers to Euclidean geometry and reads as follows[8]:

Theorem 5 (Beckman and Quarles 1953). *Let $\mathbb{R}^n$ for $n \geq 2$ be endowed with the standard Euclidean inner product $\langle \cdot \mid \cdot \rangle$. The associated norm is given by $\|x\| := \sqrt{\langle x \mid x \rangle}$. Let δ be any fixed positive real number and $f : \mathbb{R}^n \to \mathbb{R}^n$ any map such that $\|x - y\| = \delta \Rightarrow \|f(x) - f(y)\| = \delta$; then f is a Euclidean motion, i.e., $f \in \mathbb{R}^n \rtimes \mathsf{O}(n)$.*

Note that there are three obvious points which let the result of Beckman and Quarles in Euclidean space appear somewhat stronger than the theorem of Alexandrov in Minkowski space:

1. The conclusion of Theorem 5 holds for *any* $\delta \in \mathbb{R}_+$, whereas Alexandrov's theorem singles out lightlike distances.
2. In Theorem 5, $n = 2$ is not excluded.
3. In Theorem 5, f is not required to be a bijection, so that we did not assume the existence of an inverse map f^{-1}. Correspondingly, there is no assumption that f^{-1} also preserves the distance δ.

4.2.5 The Impact of The Law of Inertia

In this subsection we wish to discuss the extent to which the law of inertia already determines the automorphism group of spacetime.

The law of inertia privileges a subset of paths in spacetime form among all paths; it defines a so-called *path structure* [16, 18]. These privileged paths correspond to the motions of privileged objects called *free particles*. The existence of such privileged objects is by no means obvious and must be taken as a contingent and particularly kind property of nature. It has been known for long [35, 44, 45] how to operationally construct timescales and spatial reference frames relative to which free particles will move uniformly and on straight lines respectively – all of them! (A summary of these papers is given in [25].) These special timescales and spatial reference frames were termed *inertial* by Ludwig Lange [35]. Their existence must again be taken as

[8] In fact, Beckman and Quarles proved the conclusion of Theorem 5 under slightly weaker hypotheses: They allowed the map f to be "many-valued", that is, to be a map $f : \mathbb{R}^n \to \mathcal{S}^n$, where $\mathcal{S}^n$ is the set of non-empty subsets of $\mathbb{R}^n$, such that $\|x - y\| = \delta \Rightarrow \|x' - y'\| = \delta$ for any $x' \in f(x)$ and any $y' \in f(y)$. However, given the statement of Theorem 5, it is immediate that such "many-valued maps" must necessarily be single-valued. To see this, assume that $x_* \in \mathbb{R}^n$ has the two image points y_1, y_2 and define $h_i : \mathbb{R}^n \to \mathbb{R}^n$ for $i = 1, 2$ such that $h_1(x) = h_2(x) \in f(x)$ for all $x \neq x_*$ and $h_i(x_*) = y_i$. Then, according to Theorem 5, h_i must both be Euclidean motions. Since they are continuous and coincide for all $x \neq x_*$, they must also coincide at x_*.

a very particular and very kind feature of Nature. Note that "uniform in time" and "spatially straight" together translate to "straight in spacetime". We also emphasise that "straightness" of ensembles of paths can be characterised intrinsically, e.g., by the Desargues property [41]. All this is true if free particles are given. We do not discuss at this point whether and how one should characterise them independently (cf. [23]).

The spacetime structure so defined is usually referred to as *projective*. It it not quite that of an affine space, since the latter provides in addition each straight line with a distinguished two-parameter family of parametrisations, corresponding to a notion of *uniformity* with which the line is traced through. Such a privileged parametrisation of spacetime paths is not provided by the law of inertia, which only provides privileged parametrisations of spatial paths, which we already took into account in the projective structure of spacetime. Instead, an affine structure of spacetime may once more be motivated by another contingent property of Nature, shown by the existence of elementary clocks (atomic frequencies) which do define the same uniformity structure on inertial world lines – all of them! Once more this is a highly non-trivial and very kind feature of Nature. In this way we would indeed arrive at the statement that spacetime is an affine space. However, as we shall discuss in this subsection, the affine group already emerges as automorphism group of inertial structures without the introduction of elementary clocks.

First we recall the main theorem of affine geometry. For that we make the following

Definition 3. Three points in an affine space are called **collinear** iff they are contained in a single line. A map between affine spaces is called a **collineation** iff it maps each triple of collinear points to collinear points.

Note that in this definition no other condition is required of the map, like, e.g., injectivity. The main theorem now reads as follows:

Theorem 6. *A bijective collineation of a real affine space of dimension $n \geq 2$ is necessarily an affine map.*

A proof may be found in [6]. That the theorem is non-trivial can, e.g., be seen from the fact that it is not true for complex affine spaces. The crucial property of the real number field is that it does not allow for a non-trivial automorphisms (as field).

A particular consequence of Theorem 6 is that bijective collineation are necessarily continuous (in the natural topology of affine space). This is of interest for the applications we have in mind for the following reason: Consider the set P of all lines in some affine space S. P has a natural topology induced from S. Theorem 6 now implies that bijective collineations of S act as homeomorphism of P. Consider an open subset $\Omega \subset P$ and the subset of all collineations that fix Ω (as set, not necessarily its points). Then these collineations also fix the boundary $\partial\Omega$ of Ω in P. For example, if Ω is the set of all timelike lines in Minkowski space, i.e., with a slope less than some chosen value relative to some fixed direction, then it follows that the bijective collineations which together with their inverse map timelike lines to timelike lines also maps the lightcone to the lightcone. It immediately follows that it must be the composition of a Poincaré transformation as a constant dilation. Note that this argument also works in two spacetime dimensions, where the Alexandrov-Zeeman result does not hold.

The application we have in mind is to inertial motions, which are given by lines in affine space. In that respect Theorem 6 is not quite appropriate. Its hypotheses are weaker than needed, insofar as it would suffice to require straight lines to be mapped to straight lines. But, more importantly, the hypotheses are also stronger than what seems physically justifiable, insofar as not every line is realisable by an inertial motion. In particular, one would like to know whether Theorem 6 can still be derived by restricting to *slow collineations*, which one may define by the property that the corresponding lines should have a slope less than some non-zero angle (in whatever measure, as long as the set of slow lines is open in the set of all lines) from a given (time-)direction. This is indeed the case, as one may show from going through the proof of Theorem 6. Slightly easier to prove is the following:

Theorem 7. *Let F be a bijection of real n-dimensional affine space that maps slow lines to slow lines, then F is an affine map.*

A proof may be found in [26]. If "slowness" is defined via the lightcone of a Minkowski metric g, one immediately obtains the result that the affine maps must be composed from Poincaré transformations and dilations. The reason is

Lemma 1. *Let V be a finite dimensional real vector space of dimension $n \geq 2$ and g be a non-degenerate symmetric bilinear form on V of signature $(1, n-1)$. Let h be any other symmetric bilinear form on V. The 'light cones' for both forms are defined by $\mathcal{L}_g := \{v \in V \mid g(v,v) = 0\}$ and $\mathcal{L}_h := \{v \in V \mid h(v,v) = 0\}$. Suppose $\mathcal{L}_g \subseteq \mathcal{L}_h$, then $h = \alpha\, g$ for some $\alpha \in \mathbb{R}$.*

Proof. Let $\{e_0, e_1, \cdots, e_{n-1}\}$ be a basis of V such that $g_{ab} := g(e_a, e_b) = \mathrm{diag}(1, -1, \cdots, -1)$. Then $(e_0 \pm e_a) \in \mathcal{L}_g$ for $1 \leq a \leq n-1$ implies (we write $h_{ab} := h(e_a, e_b)$): $h_{0a} = 0$ and $h_{00} + h_{aa} = 0$. Further, $(\sqrt{2}e_0 + e_a + e_b) \in \mathcal{L}_g$ for $1 \leq a < b \leq n-1$ then implies $h_{ab} = 0$ for $a \neq b$. Hence $h = \alpha\, g$ with $\alpha = h_{00}$. □

This can be applied as follows: If $F : S \to S$ is affine and maps lightlike lines to lightlike lines, then the associated linear map $f : V \to V$ maps lightlike vectors to lightlike vectors. Hence $h(v,v) := g(f(v), f(v))$ vanishes if $g(v,v)$ vanishes and therefore $h = \alpha g$ by Lemma 1. Since $f(v)$ is timelike if v is timelike, α is positive. Hence we may define $f' := f/\sqrt{\alpha}$ and have $g(f'(v), f'(v)) = g(v,v)$ for all $v \in v$, saying that f' is a Lorentz transformation. f is the composition of a Lorentz transformation and a dilation by $\sqrt{\alpha}$.

4.2.6 *The Impact of Relativity*

As is well known, the two main ingredients in Special Relativity are the Principle of Relativity (henceforth abbreviated by PR) and the principle of the constancy of light.

We have seen above that, due to Alexandrov's Theorem, the latter almost suffices to arrive at the Poincaré group. In this section we wish to address the complementary question: Under what conditions and to what extent can the RP *alone* justify the Poincaré group?

This question was first addressed by Ignatowsky [30], who showed that under a certain set of technical assumptions (not consistently spelled out by him) the RP alone suffices to arrive at a spacetime symmetry group which is either the inhomogeneous Galilei or the inhomogeneous Lorentz group, the latter for some yet undetermined limiting velocity c.

More precisely, what is actually shown in this fashion is, as we will see, that the relativity group must contain either the proper orthochronous Galilei or Lorentz group, if the group is required to comprise at least spacetime translations, spatial rotations, and boosts (velocity transformations). What we hence gain is the group-theoretic insight of how these transformations must combine into a common group, given that they form a group at all. We do not learn anything about other transformations, like spacetime reflections or dilations, whose existence we neither required nor ruled out at this level.

The work of Ignatowsky was put into a logically more coherent form by Franck & Rothe [21, 22], who showed that some of the technical assumptions could be dropped. Further formal simplifications were achieved by Berzi & Gorini [7]. Below we shall basically follow their line of reasoning, except that we do not impose the continuity of the transformations as a requirement, but conclude it from their preservation of the inertial structure plus bijectivity. See also [2] for an alternative discussion on the level of Lie algebras.

For further determination of the automorphism group of spacetime we invoke the following principles:

ST1: Homogeneity of spacetime
ST2: Isotropy of space
ST3: Galilean principle of relativity

We take ST1 to mean that the sought-for group should include all translations and hence be a subgroup of the general affine group. With respect to some chosen basis, it must be of the form $\mathbb{R}^4 \rtimes \mathsf{G}$, where G is a subgroup of $\mathsf{GL}(4,\mathbb{R})$. ST2 is interpreted as saying that G should include the set of all spatial rotations. If, with respect to some frame, we write the general element $A \in \mathsf{GL}(4,\mathbb{R})$ in a $1+3$ split form (thinking of the first coordinate as time, the other three as space), we want G to include all

$$R(\mathbf{D}) = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{D} \end{pmatrix}, \qquad \text{where} \quad \mathbf{D} \in \mathsf{SO}(3)\,. \tag{4.16}$$

Finally, ST3 says that velocity transformations, henceforth called "boosts", are also contained in G. However, at this stage we do not know how boosts are to be represented mathematically. Let us make the following assumptions:

B1: Boosts $B(\boldsymbol{v})$ are labelled by a vector $\boldsymbol{v} \in B_c(\mathbb{R}^3)$, where $B_c(\mathbb{R}^3)$ is the open ball in $\mathbb{R}^3$ of radius c. The physical interpretation of $\boldsymbol{v}$ shall be that of the boost velocity, as measured in the system from which the transformation is carried out. We allow c to be finite or infinite ($B_\infty(\mathbb{R}^3) = \mathbb{R}^3$). $\boldsymbol{v} = \mathbf{0}$ corresponds to the identity transformation, i.e. $B(\mathbf{0}) = \mathrm{id}_{\mathbb{R}^4}$. We also assume that $\boldsymbol{v}$, considered as coordinate function on the group, is continuous.

B2: As part of ST2 we require equivariance of boosts under rotations:

$$R(\mathbf{D}) \cdot B(\boldsymbol{v}) \cdot R(\mathbf{D}^{-1}) = B(\mathbf{D} \cdot \boldsymbol{v})\,. \tag{4.17}$$

The latter assumption allows us to restrict attention to boost in a fixed direction, say that of the positive x-axis. Once their analytical form is determined as function of v, where $\boldsymbol{v} = v\boldsymbol{e}_x$, we deduce the general expression for boosts using (4.17) and (4.16). We make no assumptions involving space reflections.[9] We now restrict attention to $\boldsymbol{v} = v\boldsymbol{e}_x$. We wish to determine the most general form of $B(\boldsymbol{v})$ compatible with all requirements put so far. We proceed in several steps:

1. Using an arbitrary rotation $\mathbf{D}$ around the x-axis, so that $\mathbf{D} \cdot \boldsymbol{v} = \boldsymbol{v}$, Eq. (4.17) allows to prove that

$$B(v\boldsymbol{e}_x) = \begin{pmatrix} \mathbf{A}(v) & 0 \\ 0 & \alpha(v)\mathbf{1}_2 \end{pmatrix}, \tag{4.18}$$

where here we wrote the 4×4 matrix in a $2 + 2$ decomposed form. (i.e., $\mathbf{A}(v)$ is a 2×2 matrix and $\mathbf{1}_2$ is the 2×2 unit-matrix). Applying (4.17) once more, this time using a π-rotation about the y-axis, we learn that α is an even function, i.e.,

$$\alpha(v) = \alpha(-v)\,. \tag{4.19}$$

Below we will see that $\alpha(v) \equiv 1$.

2. Let us now focus on $\mathbf{A}(v)$, which defines the action of the boost in the $t-x$ plane. We write

$$\begin{pmatrix} t \\ x \end{pmatrix} \mapsto \begin{pmatrix} t' \\ x' \end{pmatrix} = \mathbf{A}(v) \cdot \begin{pmatrix} t \\ x \end{pmatrix} = \begin{pmatrix} a(v) & b(v) \\ c(v) & d(v) \end{pmatrix} \cdot \begin{pmatrix} t \\ x \end{pmatrix}. \tag{4.20}$$

We refer to the system with coordinates (t, x) as K and that with coordinates (t', x') as K'. From (4.20) and the inverse (which is elementary to compute) one infers that the velocity v of K' with respect to K and the velocity v' of K with respect to K' are given by

[9] Some derivations in the literature of the Lorentz group do not state the equivariance property (4.17) explicitly, though they all use it (implicitly), usually in statements to the effect that it is sufficient to consider boosts in one fixed direction. Once this restriction is effected, a one-dimensional spatial reflection transformation is considered to relate a boost transformation to that with opposite velocity. This then gives the impression that reflection equivariance is also invoked, though this is not necessary in spacetime dimensions greater than two, for (4.17) allows to invert one axis through a 180° rotation about a perpendicular one.

$$v \;=-\; c(v)/d(v)\,, \tag{4.21a}$$

$$v' \;=-\; v\,d(v)/a(v) \;=:\; \varphi(v)\,. \tag{4.21b}$$

Since the transformation $K' \to K$ is the inverse of $K \to K'$, the function $\varphi : (-c, c) \to (-c, c)$ obeys

$$\mathbf{A}(\varphi(v)) = (\mathbf{A}(v))^{-1}\,. \tag{4.22}$$

Hence φ is a bijection of the open interval $(-c, c)$ onto itself and obeys

$$\varphi \circ \varphi = \mathrm{id}_{(-c,c)}\,. \tag{4.23}$$

3. Next we determine φ. Once more using (4.17), where $\mathbf{D}$ is a π-rotation about the y-axis, shows that the functions a and d in (4.18) are even and the functions b and c are odd. The definition (4.21b) of φ then implies that φ is odd. Since we assumed $\boldsymbol{v}$ to be a continuous coordinatisation of a topological group, the map φ must also be continuous (since the inversion map, $g \mapsto g^{-1}$, is continuous in a topological group). A standard theorem now states that a continuous bijection of an interval of $\mathbb{R}$ onto itself must be strictly monotonic. Together with (4.23) this implies that φ is either the identity or minus the identity map.[10] If it is the identity map, evaluation of (4.22) shows that either the determinant of $\mathbf{A}(v)$ must equals -1, or that $\mathbf{A}(v)$ is the identity for all $\boldsymbol{v}$. We exclude the second possibility straightaway and the first one on the grounds that we required $\mathbf{A}(v)$ be the identity for $v = 0$. Also, in that case, (4.22) implies $A^2(v) = \mathrm{id}$ for all $v \in (-c, c)$. We conclude that $\varphi = -\mathrm{id}$, which implies that the relative velocity of K with respect to K' is minus the relative velocity of K' with respect to K. Plausible as it might seem, there is no a priori reason why this should be so.[11] On the face of it, the RP only implies (4.23), not the stronger relation $\varphi(v) = -v$. This was first pointed out in [7].
4. We briefly revisit (4.19). Since we have seen that $B(-v\boldsymbol{e}_x)$ is the inverse of $B(v\boldsymbol{e}_x)$, we must have $\alpha(-v) = 1/\alpha(v)$, so that (4.19) implies $\alpha(v) \equiv \pm 1$. But only $\alpha(v) \equiv +1$ is compatible with our requirement that $B(\mathbf{0})$ be the identity.
5. Now we return to the determination of $\mathbf{A}(v)$. Using (4.21) and $\varphi = -\mathrm{id}$, we write

$$\mathbf{A}(v) = \begin{pmatrix} a(v) & b(v) \\ -va(v) & a(v) \end{pmatrix} \tag{4.24}$$

---

[10] The simple proof is as follows, where we write $v' := \varphi(v)$ to save notation, so that (4.23) now reads $v'' = v$. First assume that φ is strictly monotonically increasing, then $v' > v$ implies $v = v'' > v'$, a contradiction, and $v' < v$ implies $v = v'' < v'$, likewise a contradiction. Hence $\varphi = \mathrm{id}$ in this case. Next assume φ is strictly monotonically decreasing. Then $\tilde{\varphi} := -\varphi$ is a strictly monotonically increasing map of the interval $(-c, c)$ to itself that obeys (4.23). Hence, as just seen, $\tilde{\varphi} = \mathrm{id}$, i.e., $\varphi = -\mathrm{id}$.

[11] Note that v and v' are measured with different sets of rods and clocks.

and

$$\Delta(v) := \det\bigl(\mathbf{A}(v)\bigr) = a(v)\bigl[a(v) + vb(v)\bigr]. \tag{4.25}$$

Equation $\mathbf{A}(-v) = (\mathbf{A}(v))^{-1}$ is now equivalent to

$$a(-v) = a(v)/\Delta(v), \tag{4.26a}$$
$$b(-v) = -b(v)/\Delta(v). \tag{4.26b}$$

Since, as already seen, a is an even and b is an odd function, (4.26) is equivalent to $\Delta(v) \equiv 1$, i.e. the unimodularity of $B(\boldsymbol{v})$. Equation (4.25) then allows to express b in terms of a:

$$b(v) = \frac{a(v)}{v}\left[\frac{1}{a^2(v)} - 1\right]. \tag{4.27}$$

6. Our problem is now reduced to the determination of the single function a. This we achieve by employing the requirement that the composition of two boosts in the same direction results again in a boost in that direction, i.e.,

$$\mathbf{A}(v) \cdot \mathbf{A}(v') = \mathbf{A}(v''). \tag{4.28}$$

According to (4.24) each matrix $\mathbf{A}(v)$ has equal diagonal entries. Applied to the product matrix on the left hand side of (4.28) this implies that $v^{-2}(a^{-2}(v) - 1)$ is independent of v, i.e., equal to some constant k whose physical dimension is that of an inverse velocity squared. Hence we have

$$a(v) = \frac{1}{\sqrt{1 + kv^2}}, \tag{4.29}$$

where we have chosen the positive square root since we require $a(0) = 1$. The other implications of (4.28) are

$$a(v)a(v')(1 - kvv') = a(v''), \tag{4.30a}$$
$$a(v)a(v')(1 + vv') = v''a(v''), \tag{4.30b}$$

from which we deduce

$$v'' = \frac{v + v'}{1 - kvv'}. \tag{4.31}$$

Conversely, (4.29) and (4.31) imply (4.30). We conclude that (4.28) is equivalent to (4.29) and (4.31).

7. So far a boost in x direction has been shown to act non-trivially only in the $t - x$ plane, where its action is given by the matrix that results from inserting (4.27) and (4.29) into (4.24):

$$\mathbf{A}(v) = \begin{pmatrix} a(v) & kv\,a(v) \\ -v\,a(v) & a(v), \end{pmatrix} \qquad \text{where} \quad a(v) = 1/\sqrt{1 + kv^2}. \tag{4.32}$$

- If $k > 0$ we rescale $t \mapsto \tau := t/\sqrt{k}$ and set $\sqrt{k}\, v := \tan\alpha$. Then (4.32) is seen to be a Euclidean rotation with angle α in the $\tau - x$ plane. The velocity spectrum is the whole real line plus infinity, i.e., a circle, corresponding to $\alpha \in [0, 2\pi]$, where 0 and 2π are identified. Accordingly, the composition law (4.31) is just ordinary addition for the angle α. This causes several paradoxa when v is interpreted as velocity. For example, composing two finite velocities v, v' which satisfy $vv' = 1/k$ results in $v'' = \infty$, and composing two finite and positive velocities, each of which is greater than $1/\sqrt{k}$, results in a finite but negative velocity. In this way the successive composition of finite positive velocities could also result in zero velocity. The group $\mathsf{G} \subset \mathsf{GL}(n, \mathbb{R})$ obtained in this fashion is, in fact, $\mathsf{SO}(4)$. This group may be uniquely characterised as the largest connected group of bijections of $\mathbb{R}^4$ that preserves the Euclidean distance measure. In particular, it treats time symmetrically with all space directions, so that no invariant notion of time-orientability can be given in this case.
- For $k = 0$ the transformations are just the ordinary boosts of the Galilei group. The velocity spectrum is the whole real line (i.e., v is unbounded but finite) and G is the Galilei group. The law for composing velocities is just ordinary vector addition.
- Finally, for $k < 0$, one infers from (4.31) that $c := 1/\sqrt{-k}$ is an upper bound for all velocities, in the sense that composing two velocities taken from the interval $(-c, c)$ always results in a velocity from within that interval. Writing $\tau := ct$, $v/c =: \beta =: \tanh\rho$, and $\gamma = 1/\sqrt{1-\beta^2}$, the matrix (4.32) is seen to be a *Lorentz boost* or *hyperbolic motion* in the $\tau - x$ plane:

$$\begin{pmatrix} \tau \\ x \end{pmatrix} \mapsto \begin{pmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{pmatrix} \cdot \begin{pmatrix} \tau \\ x \end{pmatrix} = \begin{pmatrix} \cosh\rho & -\sinh\rho \\ -\sinh\rho & \cosh\rho \end{pmatrix} \cdot \begin{pmatrix} \tau \\ x \end{pmatrix} . \tag{4.33}$$

  The quantity

$$\rho := \tanh^{-1}(v/c) = \tanh^{-1}(\beta) \tag{4.34}$$

  is called *rapidity*[12]. If rewritten in terms of the corresponding rapidities the composition law (4.31) reduces to ordinary addition: $\rho'' = \rho + \rho'$.

This shows that only the Galilei and the Lorentz group survive as candidates for any symmetry group implementing the RP. Once the Lorentz group for velocity parameter c is chosen, one may fully characterise it by its property to leave a certain symmetric bilinear form invariant. In this sense we geometric structure of Minkowski space can be deduced. This closes the circle to where we started from in Section 4.2.3.

[12] This term was coined by Robb [43], but the quantity was used before by others; compare [47].

4.2.7 Local Versions

In the previous sections we always understood an automorphisms of a structured set (spacetime) as a bijection. Mathematically this seems an obvious requirement, but from a physical point of view this is less clear. The physical law of inertia provides us with distinguished motions *locally* in space and time. Hence one may attempt to relax the condition for structure preserving maps, so as to only preserve inertial motions *locally*. Hence we ask the following question: What are the most general maps that *locally* map segments of straight lines to segments of straight lines? This local approach has been pursued by [20].

To answer this question completely, let us (locally) identify spacetime with $\mathbb{R}^n$ where $n \geq 2$ and assume the map to be C^3, that is, three times continuously differentiable.[13] So let $U \subseteq \mathbb{R}^n$ be an open subset and determine all C^3 maps $f : U \to \mathbb{R}^n$ that map straight segments in U into straight segments in $\mathbb{R}^n$. In coordinates we write $x = (x^1, \cdots, x^n) \in U$ and $y = (y^1, \cdots, y^n) \in f(U) \subseteq \mathbb{R}^n$, so that $y^\mu := f^\mu(x)$. A straight segment in U is a curve $\gamma : I \to U$ (the open interval $I \subseteq \mathbb{R}$ is usually taken to contain zero) whose acceleration is pointwise proportional to its velocity. This is equivalent to saying that it can be parametrised so as to have zero acceleration, i.e., $\gamma(s) = as + b$ for some $a, b \in \mathbb{R}^n$.

For the image path $f \circ \gamma$ to be again straight its acceleration, $(f'' \circ \gamma)(a, a)$, must be proportional to its velocity, $(f' \circ \gamma)(a)$, where the factor of proportionality, C, depends on the point of the path and separately on a. Hence, in coordinates, we have

$$f^\mu_{,\lambda\sigma}(as + b)a^\lambda a^\sigma = f^\mu_{,\nu}(as + b)a^\nu \, C(as + b, a) \tag{4.35}$$

For each b this must be valid for all (a, s) in a neighbourhood of zero in $\mathbb{R}^n \times \mathbb{R}$. Taking the second derivatives with respect to a, evaluation at $a = 0$, $s = 0$ leads to

$$f^\mu_{,\lambda\sigma} = \Gamma^\nu_{\lambda\sigma} f^\mu_{,\nu} \,, \tag{4.36a}$$

where

$$\Gamma^\nu_{\lambda\sigma} := \delta^\nu_\lambda \psi_\sigma + \delta^\nu_\sigma \psi_\lambda \tag{4.36b}$$

$$\psi_\sigma := \left.\frac{\partial C(\cdot, a)}{\partial a^\sigma}\right|_{a=0} \tag{4.36c}$$

Here we suppressed the remaining argument b. Equation (4.36) is valid at each point in U. Integrability of (4.36a) requires that its further differentiation is totally symmetric with respect to all lower indices (here we use that the map f is C^3). This leads to

$$R^\mu{}_{\alpha\beta\gamma} := \partial_\beta \Gamma^\mu_{\alpha\gamma} + \Gamma^\nu_{\sigma\beta}\Gamma^\sigma_{\alpha\gamma} - (\beta \leftrightarrow \gamma) = 0 \,. \tag{4.37}$$

[13] This requirement distinguishes the present (local) from the previous (global) approaches, in which not even continuity needed to be assumed.

Inserting (4.36b) one can show (upon taking traces over $\mu\alpha$ and $\mu\gamma$) that the resulting equation is equivalent to

$$\psi_{\alpha,\beta} = \psi_\alpha \psi_\beta\,. \tag{4.38}$$

In particular $\psi_{\alpha,\beta} = \psi_{\beta,\alpha}$ so that there is a local function $\psi : U \to \mathbb{R}$ (if U is simply connected, as we shall assume) for which $\psi_\alpha = \psi_{,\alpha}$. Equation (4.38) is then equivalent to $\partial_\alpha\partial_\beta \exp(-\psi) = 0$ so that $\psi(x) = -\ln(p \cdot x + q)$ for some $p \in \mathbb{R}^n$ and $q \in \mathbb{R}$. Using $\psi_\sigma = \psi_{,\sigma}$ and (4.38), Eq. (4.36a) is equivalent to $\partial_\lambda\partial_\sigma\big[f^\mu \exp(-\psi)\big] = 0$, which finally leads to the result that the most general solution for f is given by

$$f(x) = \frac{A \cdot x + a}{p \cdot x + q}\,. \tag{4.39}$$

Here A is a $n \times n$ matrix, a and q vectors in $\mathbb{R}^n$, and $q \in \mathbb{R}$. p and q must be such that U does not intersect the hyperplane $H(p,q) := \{x \in \mathbb{R}^n \mid p \cdot x + q = 0\}$ where f becomes singular, but otherwise they are arbitrary. Iff $H(p,q) \neq \emptyset$, i.e., iff $p \neq 0$, the transformations (4.39) are not affine. In this case they are called proper projective.

Are there physical reasons to rule out such proper projective transformations? A structural argument is that they do not leave any subset of $\mathbb{R}^n$ invariant and that they hence cannot be considered as automorphism group of any subdomain. A physical argument is that two separate points that move with the same velocity cease to do so if their worldlines are transformed by by a proper projective transformation. In particular, a rigid motion of an extended body (undergoing inertial motion) ceases to be rigid if so transformed (cf. [17], p. 16). An illustrative example is the following: Consider the one-parameter (σ) family of parallel lines $x(s,\sigma) = se_0 + \sigma e_1$ (where s is the parameter along each line), and the proper projective map $f(x) = x/(-e_0 \cdot x + 1)$ which becomes singular on the hyperplane $x^0 = 1$. The one-parameter family of image lines

$$y(s,\sigma) := f\big(x(s,\sigma)\big) = \frac{se_0 + \sigma e_1}{1 - s} \tag{4.40}$$

have velocities

$$\partial_s y(s,\sigma) = \frac{qe_0 + \sigma e_1}{(1-s)^2} \tag{4.41}$$

whose directions are independent of s, showing that they are indeed straight. However, the velocity directions now depend on σ, showing that they are not parallel anymore.

Let us, regardless of this, for the moment take seriously the transformations (4.39). One may reduce them to the following form of generalised boosts, discarding translations and rotations and using equivariance with respect to the latter (we restrict to four spacetime dimensions from now on):

$$t' = \frac{a(v)t + b(v)(v \cdot \boldsymbol{x})}{A(v) + B(v)t + D(v)(\boldsymbol{v} \cdot \boldsymbol{x})}\,, \tag{4.42a}$$

$$\boldsymbol{x}'_{\parallel} \;=\; \frac{d(v)\boldsymbol{v}t + e(v)\boldsymbol{x}_{\parallel}}{A(v) + B(v)t + D(v)(\boldsymbol{v}\cdot\boldsymbol{x})}\,, \tag{4.42b}$$

$$\boldsymbol{x}'_{\perp} \;=\; \frac{f(v)\boldsymbol{x}_{\perp}}{A(v) + B(v)t + D(v)(\boldsymbol{v}\cdot\boldsymbol{x})}\,. \tag{4.42c}$$

where $\boldsymbol{v} \in \mathbb{R}^3$ represents the boost velocity, $v := \|\boldsymbol{v}\|$ its modulus, and all functions of v are even. The subscripts $\parallel$ and $\perp$ refer to the components parallel and perpendicular to $\boldsymbol{v}$. Now one imposes the following conditions which allow to determine the eight functions a, b, d, e, f, A, B, D, of which only seven are considered independent since common factors of the numerator and denominator cancel (we essentially follow [38]):

1. The origin $\boldsymbol{x}' = 0$ has velocity $\boldsymbol{v}$ in the unprimed coordinates, leading to $e(v) = -d(v)$ and thereby eliminating e as independent function.
2. The origin $\boldsymbol{x} = 0$ has velocity $-\boldsymbol{v}$ in the primed coordinates, leading to $d(v) = -a(v)$ and thereby eliminating d as independent function.
3. Reciprocity: The transformation parametrised by $-\boldsymbol{v}$ is the inverse of that parametrised by $\boldsymbol{v}$, leading to relations $A = A(a, b, v)$, $B = B(D, a, b, v)$, and $f = A$, thereby eliminating A, B, f as independent functions. Of the remaining three functions a, b, D an overall factor in the numerator and denominator can be split off so that two free functions remain.
4. Transitivity: The composition of two transformations of the type (4.42) with parameters $\boldsymbol{v}$ and $\boldsymbol{v}'$ must be again of this form with some parameter $\boldsymbol{v}''(\boldsymbol{v}, \boldsymbol{v}')$, which turns out to be the same function of the velocities $\boldsymbol{v}$ and $\boldsymbol{v}'$ as in Special Relativity (Einstein's addition law), for reasons to become clear soon. This allows to determine the last two functions in terms of two constants c and R whose physical dimensions are that of a velocity and of a length respectively. Writing, as usual, $\gamma(v) := 1/\sqrt{1 - v^2/c^2}$ the final form is given by

$$t' \;=\; \frac{\gamma(v)(t - \boldsymbol{v}\cdot\boldsymbol{x}/c^2)}{1 - \big(\gamma(v) - 1\big)ct/R + \gamma(v)\boldsymbol{v}\cdot\boldsymbol{x}/Rc}\,, \tag{4.43a}$$

$$\boldsymbol{x}'_{\parallel} \;=\; \frac{\gamma(v)(\boldsymbol{x}_{\parallel} - \boldsymbol{v}t)}{1 - \big(\gamma(v) - 1\big)ct/R + \gamma(v)\boldsymbol{v}\cdot\boldsymbol{x}/Rc}\,, \tag{4.43b}$$

$$\boldsymbol{x}'_{\perp} \;=\; \frac{\boldsymbol{x}_{\perp}}{1 - \big(\gamma(v) - 1\big)ct/R + \gamma(v)\boldsymbol{v}\cdot\boldsymbol{x}/Rc}\,. \tag{4.43c}$$

In the limit as $R \to \infty$ this approaches an ordinary Lorentz boost:

$$L(\boldsymbol{v}) : (t, \boldsymbol{x}_{\parallel}, \boldsymbol{x}_{\perp}) \mapsto \big(\gamma(v)(t - \boldsymbol{v}\cdot\boldsymbol{x}/c^2)\,,\, \gamma(v)(\boldsymbol{x}_{\parallel} - \boldsymbol{v}t)\,,\, \boldsymbol{x}_{\perp}\big)\,. \tag{4.44}$$

Moreover, for finite R the map (4.43) is conjugate to (4.44) with respect to a time dependent deformation. To see this, observe that the common denominator in (4.43) is just $(R + ct)/(R + ct')$, whereas the numerators correspond to (4.44). Hence, introducing the deformation map

$$\phi : (t, \boldsymbol{x}) \mapsto \left(\frac{t}{1 - ct/R}, \frac{\boldsymbol{x}}{1 - ct/R} \right) \tag{4.45}$$

and denoting the map $(t, \boldsymbol{x}) \mapsto (t', \boldsymbol{x}')$ in (4.43) by f, we have

$$f = \phi \circ L(\boldsymbol{v}) \circ \phi^{-1} . \tag{4.46}$$

Note that ϕ is singular at the hyperplane $t = R/c$ and has no point of the hyperplane $t = -R/c$ in its image. The latter hyperplane is the singularity set of ϕ^{-1}. Outside the hyperplanes $t = \pm R/c$ the map ϕ relates the following time slabs in a diffeomorphic fashion:

$$0 \le t < R/c \quad \mapsto \quad 0 \le t < \infty , \tag{4.47a}$$

$$R/c < t < \infty \quad \mapsto \quad -\infty < t < -R/c , \tag{4.47b}$$

$$-\infty < t \le 0 \quad \mapsto \quad -R/c < t \le 0 . \tag{4.47c}$$

Since boosts leave the upper-half spacetime, $t > 0$, invariant (as set), (4.47a) shows that f just squashes the linear action of boosts in $0 < t < \infty$ into a non-linear action within $0 < t < R/c$, where R now corresponds to an invariant scale. Interestingly, this is the same deformation of boosts that have been recently considered in what is sometimes called *Double Special Relativity* (because there are now two, rather than just one, invariant scales, R and c), albeit there the deformation of boosts take place in momentum space where R then corresponds to an invariant energy scale; see [37] and also [32].

4.3 Selected Structures in Minkowski Space

In this section we wish to discuss in more detail some of the non-trivial structures in Minkowski. I have chosen them so as to emphasise the difference to the corresponding structures in Galilean spacetime, and also because they do not seem to be much discussed in other standard sources.

4.3.1 Simultaneity

Let us start right away by characterising those vectors for which we have an inverted Cauchy-Schwarz inequality:

Lemma 2. *Let V be of dimension $n > 2$ and $v \in V$ be some non-zero vector. The strict inverted Cauchy-Schwarz inequality,*

$$v^2 w^2 < (v \cdot w)^2 , \tag{4.48}$$

holds for all $w \in V$ linearly independent of v iff v is timelike.

Proof. Obviously v cannot be spacelike, for then we would violate (4.48) with any spacelike w. If v is lightlike then w violates (4.48) iff it is in the set $v^\perp - \text{span}\{v\}$, which is non-empty iff $n > 2$. Hence v cannot be lightlike if $n > 2$. If v is timelike we decompose $w = av + w'$ with $w' \in v^\perp$ so that $w'^2 \leq 0$, with equality iff v and w are linearly dependent. Hence

$$(v \cdot w)^2 - v^2 w^2 = -v^2\, w'^2 \geq 0\,, \tag{4.49}$$

with equality iff v and w are linearly dependent.

The next Lemma deals with the intersection of a causal line with a light cone, a situation depicted in Fig. 4.1.

Lemma 3. *Let $\mathcal{L}_p$ be the light-doublecone with vertex p and $\ell := \{r + \lambda v \mid r \in \mathbb{R}\}$ be a non-spacelike line, i.e. $v^2 \geq 0$, through $r \notin \mathcal{L}_p$. If v is timelike $\ell \cap \mathcal{L}_p$ consists of two points. If v is lightlike this intersection consists of one point if $p - r \notin v^\perp$ and is empty if $p - r \in v^\perp$. Note that the latter two statements are independent of the choice of $r \in \ell$—as they must be—, i.e. are invariant under $r \mapsto r' := r + \sigma v$, where $\sigma \in \mathbb{R}$.*

Proof. We have $r + \lambda v \in \mathcal{L}_p$ iff

$$(r + \lambda v - p)^2 = 0 \iff \lambda^2 v^2 + 2\lambda v \cdot (r - p) + (r - p)^2 = 0\,. \tag{4.50}$$

For v timelike we have $v^2 > 0$ and (4.50) has two solutions

$$\lambda_{1,2} = \frac{1}{v^2}\left\{ -v \cdot (r - p) \pm \sqrt{\big(v \cdot (r - p)\big)^2 - v^2 (r - p)^2} \right\}\,. \tag{4.51}$$

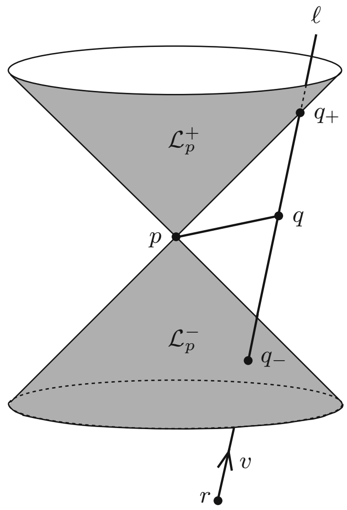


Fig. 4.1 A timelike line $\ell = \{r + \lambda v \mid \lambda \in \mathbb{R}\}$ intersects the light-cone with vertex $p \notin \ell$ in two points: q_+, its intersection with the future light-cone and q_-, its intersection with past the light cone. q is a point in between q_+ and q_-

Indeed, since $r \notin \mathcal{L}_p$, the vectors v and $r-p$ cannot be linearly dependent so that Lemma 2 implies the positivity of the expression under the square root. If v is lightlike (4.50) becomes a linear equation which is has one solution if $v\cdot(r-p) \neq 0$ and no solution if $v\cdot(r-p)=0$ [note that $(r-p)^2 \neq 0$ since $q \notin \mathcal{L}_p$ by hypothesis].

Proposition 1. *Let ℓ and $\mathcal{L}_p$ as in Lemma 3 with v timelike. Let q_+ and q_- be the two intersection points of ℓ with $\mathcal{L}_p$ and $q \in \ell$ a point between them. Then*

$$\|q-p\|_g^2 = \|q_+ - q\|_g \, \|q - q_-\|_g \, . \tag{4.52}$$

Moreover, $\|q_+ - q\|_g = \|q-q_-\|_g$ iff $p-q$ is perpendicular to v.

Proof. The vectors $(q_+ - p) = (q-p)+(q_+ - q)$ and $(q_- - p) = (q-p)+(q_- - q)$ are lightlike, which gives (note that $q-p$ is spacelike):

$$\|q-p\|_g^2 \;=\; -(q-p)^2 \;=\; (q_+ - q)^2 + 2(q-p)\cdot(q_+ - q)\,, \tag{4.53a}$$

$$\|q-p\|_g^2 \;=\; -(q-p)^2 \;=\; (q_- - q)^2 + 2(q-p)\cdot(q_- - q)\,. \tag{4.53b}$$

Since $q_+ - q$ and $q - q_-$ are parallel we have $q_+ - q = \lambda(q-q_-)$ with $\lambda \in \mathbb{R}_+$ so that $(q_+ - q)^2 = \lambda\|q_+ - q\|_g\|q-q_-\|_g$ and $\lambda(q_- - q)^2 = \|q_+ - q\|_g\|q-q_-\|_g$. Now, multiplying (4.53b) with λ and adding this to (4.53a) immediately yields

$$(1+\lambda)\,\|q-p\|_g^2 = (1+\lambda)\;\|q_+ - q\|_g\|q-q_-\|_g \, . \tag{4.54}$$

Since $1+\lambda \neq 0$ this implies (4.52). Finally, since $q_+ - q$ and $q_- - q$ are antiparallel, $\|q_+ - q\|_g = \|q_- - q\|_g$ iff $(q_+ - q) = -(q_- - q)$. Equations (4.53) now show that this is the case iff $(q-p)\cdot(q_\pm - q) = 0$, i.e., iff $(q-p)\cdot v = 0$. Hence we have shown

$$\|q_+ - q\|_g = \|q-q_-\|_g \iff (q-p)\cdot v = 0\,. \tag{4.55}$$

In other words, q is the midpoint of the segment $\overline{q_+q_-}$ iff the line through p and q is perpendicular (wrt. g) to ℓ.

The somewhat surprising feature of the first statement of this proposition is that (4.52) holds for *any* point of the segment $\overline{q_+q_-}$, not just the midpoint, as it would have to be the case for the corresponding statement in Euclidean geometry.

The second statement of Proposition 1 gives a convenient geometric characterisation of Einstein-simultaneity. Recall that an event q on a timelike line ℓ (representing an inertial observer) is defined to be Einstein-simultaneous with an event p in spacetime iff q bisects the segment $\overline{q_+q_-}$ between the intersection points q_+, q_- of ℓ with the double-lightcone at p. Hence Proposition 1 implies

Corollary 1. *Einstein simultaneity with respect to a timelike line ℓ is an equivalence relation on spacetime, the equivalence classes of which are the spacelike hyperplanes orthogonal (wrt. g) to ℓ.*

The first statement simply follows from the fact that the family of parallel hyperplanes orthogonal to ℓ form a partition (cf. Appendix 1) of spacetime.

From now on we shall use the terms "timelike line" and "inertial observer" synonymously. Note that Einstein simultaneity is only defined relative to an inertial observer. Given two inertial observers,

$$\ell \;=\; \{r + \lambda v \mid \lambda \in \mathbb{R}\} \qquad \text{first observer}\,, \tag{4.56a}$$

$$\ell' \;=\; \{r' + \lambda' v' \mid \lambda' \in \mathbb{R}\} \qquad \text{second observer}\,, \tag{4.56b}$$

we call the corresponding Einstein-simultaneity relations *ℓ-simultaneity* and *ℓ'-simultaneity*. Obviously they coincide iff ℓ and ℓ' are parallel (v and v' are linearly dependent). In this case $q' \in \ell'$ is ℓ-simultaneous to $q \in \ell$ iff $q \in \ell$ is ℓ'-simultaneous to $q' \in \ell'$. If ℓ and $\ell'$are not parallel (skew or intersecting in one point) it is generally not true that if $q' \in \ell'$ is ℓ-simultaneous to $q \in \ell$ then $q \in \ell$ is also ℓ'-simultaneous to $q' \in \ell'$. In fact, we have

Proposition 2. *Let ℓ and ℓ' two non-parallel timelike likes. There exists a unique pair $(q, q') \in \ell \times \ell'$ so that q' is ℓ-simultaneous to q and q is ℓ' simultaneous to q'.*

Proof. We parameterise ℓ and ℓ' as in (4.56). The two conditions for q' being ℓ-simultaneous to q and q being ℓ'-simultaneous to q' are $(q-q')\cdot v = 0 = (q-q')\cdot v'$. Writing $q = r + \lambda v$ and $q' = r' + \lambda' v'$ this takes the form of the following matrix equation for the two unknowns λ and λ':

$$\begin{pmatrix} v^2 & -v \cdot v' \\ v \cdot v' & -v'^2 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda' \end{pmatrix} = \begin{pmatrix} (r'-r)\cdot v \\ (r'-r)\cdot v' \end{pmatrix}. \tag{4.57}$$

This has a unique solution pair (λ, λ'), since for linearly independent timelike vectors v and v' Lemma 2 implies $(v \cdot v')^2 - v^2 v'^2 > 0$. Note that if ℓ and ℓ' intersect $q = q' =$ intersection point.

Clearly, Einstein-simultaneity is conventional and physics proper should not depend on it. For example, the fringe-shift in the Michelson-Morley experiment is independent of how we choose to synchronise clocks. In fact, it does not even make use of any clock. So what is the general definition of a "simultaneity structure"? It seems obvious that it should be a relation on spacetime that is at least symmetric (each event should be simultaneous to itself). Going from one-way simultaneity to the mutual synchronisation of two clocks, one might like to also require reflexivity (if p is simultaneous to q then q is simultaneous to p), though this is not strictly required in order to one-way synchronise each clock in a set of clocks with one preferred "master clock", which is sufficient for many applications.

Moreover, if we like to speak of the mutual simultaneity of sets of more than two events we need an equivalence relation on spacetime. The equivalence relation should be such that each inertial observer intersect each equivalence class precisely once. Let us call such a simultaneity structure "admissible". Clearly there are zillions of such structures: just partition spacetime into any set of appropriate[14] spacelike hypersurfaces (there are more possibilities at this point, like families of forward or backward lightcones). An *absolute* admissible simultaneity structure would be one which is invariant (cf. Appendix 1) under the automorphism group of spacetime. We have

Proposition 3. *There exits precisely one admissible simultaneity structure which is invariant under the inhomogeneous proper orthochronous Galilei group and none that is invariant under the inhomogeneous proper orthochronous Lorentz group.*

A proof is given in [24]. There is a group-theoretic reason that highlights this existential difference:

Proposition 4. *Let G be a group with transitive action on a set S. Let* $\mathrm{Stab}(p) \subset G$ *be the stabiliser subgroup for $p \in S$ (due to transitivity all stabiliser subgroups are conjugate). Then S admits a G-invariant equivalence relation $R \subset S \times S$ iff* $\mathrm{Stab}(p)$ *is* not *maximal, that is, iff* $\mathrm{Stab}(p)$ *is properly contained in a proper subgroup H of G:* $\mathrm{Stab}(p) \subsetneq H \subsetneq G$.

A proof of this may be found in [31] (Theorem 1.12). Regarding the action of the inhomogeneous Galilei and Lorentz groups on spacetime, their stabilisers are the corresponding homogeneous groups. Now, the homogeneous Lorentz group is maximal in the inhomogeneous one, whereas the homogeneous Galilei group is not maximal in the inhomogeneous one, since it can still be supplemented by time translations without the need to also invoke space translations.[15] This, according to Proposition 4, is the group theoretic origin of the absence of any invariant simultaneity structure in the Lorentzian case.

However, one may ask whether there are simultaneity structures *relative* to some *additional* structure X. As additional structure, X, one could, for example, take an inertial reference frame, which is characterised by a foliation of spacetime by parallel timelike lines. The stabiliser subgroup of that structure within the proper orthochronous Poincaré group is given by the semidirect product of spacetime translations with all rotations in the hypersurfaces perpendicular to the lines in X:

$$\mathrm{Stab}_X(\mathsf{ILor}_{\uparrow +}) \cong \mathbb{R}^4 \rtimes \mathsf{SO}(3)\,. \tag{4.58}$$

Here the $\mathsf{SO}(3)$ only acts on the spatial translations, so that the group is also isomorphic to $\mathbb{R} \times \mathsf{E}(3)$, where $\mathsf{E}(3)$ is the group of Euclidean motions in three-dimensions (the hyperplanes perpendicular to the lines in X). We can now ask: how many admissible $\mathrm{Stab}_X(\mathsf{ILor}_{\uparrow +})$ – invariant equivalence relations are there. The answer is

[14] For example, the hypersurfaces should not be asymptotically hyperboloidal, for then a constantly accelerated observer would not intersect all of them.

[15] The homogeneous Galilei group only acts on the spatial translations, not the time translations, whereas the homogeneous Lorentz group acts irreducibly on the vector space of translations.

Proposition 5. *There exits precisely one admissible simultaneity structure which is invariant under* $\text{Stab}_X(\mathsf{ILor}_{\uparrow+})$*, where* X *represents am inertial reference frame (a foliation of spacetime by parallel timelike lines). It is given by Einstein simultaneity, that is, the equivalence classes are the hyperplanes perpendicular to the lines in* X.

The proof is given in [24]. Note again the connection to quoted group-theoretic result: The stabiliser subgroup of a point in $\text{Stab}_X(\mathsf{ILor}_{\uparrow+})$ is $\mathsf{SO}(3)$, which is clearly not maximal in $\text{Stab}_X(\mathsf{ILor}_{\uparrow+})$ since it is a proper subgroup of $\mathsf{E}(3)$ which, in turn, is a proper subgroup of $\text{Stab}_X(\mathsf{ILor}_{\uparrow+})$.

4.3.2 The Lattices of Causally and Chronologically Complete Sets

Here we wish to briefly discuss another important structure associated with causality relations in Minkowski space, which plays a fundamental rôle in modern Quantum Field Theory (see e.g., [27]). Let S_1 and S_2 be subsets of $\mathbb{M}^n$. We say that S_1 and S_2 are *causally disjoint* or *spacelike separated* iff $p_1 - p_2$ is spacelike, i.e., $(p_1 - p_2)^2 < 0$, for any $p_1 \in S_1$ and $p_2 \in S_2$. Note that because a point is not spacelike separated from itself, causally disjoint sets are necessarily disjoint in the ordinary set-theoretic sense – the converse being of course not true.

For any subset $S \subseteq \mathbb{M}^n$ we denote by S' the largest subset of $\mathbb{M}^n$ which is causally disjoint to S. The set S' is called the *causal complement* of S. The procedure of taking the causal complement can be iterated and we set $S'' := (S')'$ etc. S'' is called the *causal completion* of S. It also follows straight from the definition that $S_1 \subseteq S_2$ implies $S_1' \supseteq S_2'$ and also $S'' \supseteq S$. If $S'' = S$ we call S *causally complete*. We note that the causal complement S' of any given S is automatically causally complete. Indeed, from $S'' \supseteq S$ we obtain $(S')'' \subseteq S'$, but the first inclusion applied to S' instead of S leads to $(S')'' \supseteq S'$, showing $(S')'' = S'$. Note also that for any subset S its causal completion, S'', is the smallest causally complete subset containing S, for if $S \subseteq K \subseteq S''$ with $K'' = K$, we derive from the first inclusion by taking $''$ that $S'' \subseteq K$, so that the second inclusion yields $K = S''$. Trivial examples of causally complete subsets of $\mathbb{M}^n$ are the empty set, single points, and the total set $\mathbb{M}^n$. Others are the open diamond-shaped regions (4.15) as well as their closed counterparts:

$$\bar{U}(p,q) := (\bar{\mathcal{C}}_p^+ \cap \bar{\mathcal{C}}_q^-) \cup (\bar{\mathcal{C}}_q^+ \cap \bar{\mathcal{C}}_p^-)\,. \tag{4.59}$$

We now focus attention to the set $\text{Caus}(\mathbb{M}^n)$ of causally complete subsets of $\mathbb{M}^n$, including the empty set, $\emptyset$, and the total set, $\mathbb{M}^n$, which are mutually causally complementary. It is partially ordered by ordinary set-theoretic inclusion ($\subseteq$) (cf. Appendix 1) and carries the "dashing operation" ($'$) of taking the causal complement. Moreover, on $\text{Caus}(\mathbb{M}^n)$ we can define the operations of "meet" and "join", denoted by $\wedge$ and $\vee$ respectively, as follows: Let $S_i \in \text{Caus}(\mathbb{M}^n)$ where $i = 1, 2$, then $S_1 \wedge S_2$ is the largest causally complete subset in the intersection $S_1 \cap S_2$ and $S_1 \vee S_2$ is the smallest causally complete set containing the union $S_1 \cup S_2$.

The operations of $\wedge$ and $\vee$ can be characterised in terms of the ordinary set-theoretic intersection $\cap$ together with the dashing-operation. To see this, consider two causally complete sets, S_i where $i = 1,\ 2$, and note that the set of points that are spacelike separated from S_1 and S_2 are obviously given by $S'_1 \cap S'_2$, but also by $(S_1 \cup S_2)'$, so that

$$S'_1 \cap S'_2 \;=\; (S_1 \cup S_2)'\,, \tag{4.60a}$$

$$S_1 \cap S_2 \;=\; (S'_1 \cup S'_2)'\,. \tag{4.60b}$$

Here (4.60a) and (4.60b) are equivalent since any $S_i \in \mathrm{Caus}(\mathbb{M}^n)$ can be written as $S_i = P'_i$, namely $P_i = S'_i$. If S_i runs through all sets in $\mathrm{Caus}(\mathbb{M}^n)$ so does P_i. Hence any equation that holds generally for all $S_i \in \mathrm{Caus}(\mathbb{M}^n)$ remains valid if the S_i are replaced by S'_i.

Equation (4.60b) immediately shows that $S_1 \cap S_2$ is causally complete (since it is the $'$ of something). Taking the causal complement of (4.60a) we obtain the desired relation for $S_1 \vee S_2 := (S_1 \cup S_2)''$. Together we have

$$S_1 \wedge S_2 \;=\; S_1 \cap S_2\,, \tag{4.61a}$$

$$S_1 \vee S_2 \;=\; (S'_1 \cap S'_2)'\,. \tag{4.61b}$$

From these we immediately derive

$$(S_1 \wedge S_2)' \;=\; S'_1 \vee S'_2\,, \tag{4.62a}$$

$$(S_1 \vee S_2)' \;=\; S'_1 \wedge S'_2\,. \tag{4.62b}$$

All what we have said so far for the set $\mathrm{Caus}(\mathbb{M}^n)$ could be repeated verbatim for the set $\mathrm{Chron}(\mathbb{M}^n)$ of *chronologically complete* subsets. We say that S_1 and S_2 are *chronologically disjoint* or *non-timelike separated*, iff $S_1 \cap S_2 = \emptyset$ and $(p_1 - p_2)^2 \leq 0$ for any $p_1 \in S_1$ and $p_2 \in S_2$. S', the *chronological complement* of S, is now the largest subset of $\mathbb{M}^n$ which is chronologically disjoint to S. The only difference between the causal and the chronological complement of S is that the latter now contains lightlike separated points outside S. A set S is *chronologically complete* iff $S = S''$, where the dashing now denotes the operation of taking the chronological complement. Again, for any set S the set S' is automatically chronologically complete and S'' is the smallest chronologically complete subset containing S. Single points are chronologically complete subsets. All the formal properties regarding $'$, $\wedge$, and $\vee$ stated hitherto for $\mathrm{Caus}(\mathbb{M}^n)$ are the same for $\mathrm{Chron}(\mathbb{M}^n)$.

One major difference between $\mathrm{Caus}(\mathbb{M}^n)$ and $\mathrm{Chron}(\mathbb{M}^n)$ is that the types of diamond-shaped sets they contain are different. For example, the closed ones, (4.59), are members of both. The open ones, (4.15), are contained in $\mathrm{Caus}(\mathbb{M}^n)$ but *not* in

$\mathrm{Chron}(\mathbb{M}^n)$. Instead, $\mathrm{Chron}(\mathbb{M}^n)$ contains the closed diamonds whose 'equator'[16] have been removed. An essential structural difference between $\mathrm{Caus}(\mathbb{M}^n)$ and $\mathrm{Chron}(\mathbb{M}^n)$ will be stated below, after we have introduced the notion of a lattice to which we now turn.

To put all these formal properties into the right frame we recall the definition of a lattice. Let $(L, \leq)$ be a partially ordered set and a, b any two elements in L. Synonymously with $a \leq b$ we also write $b \geq a$ and say that a is smaller than b, b is bigger than a, or b majorises a. We also write $a < b$ if $a \leq b$ and $a \neq b$. If, with respect to $\leq$, their greatest lower and least upper bound exist, they are denoted by $a \wedge b$ – called the "meet of a and b" – and $a \vee b$ – called the "join of a and b" – respectively. A partially ordered set for which the greatest lower and least upper bound exist for any pair a, b of elements from L is called a *lattice*.

We now list some of the most relevant additional structural elements lattices can have: A lattice is called *complete* if greatest lower and least upper bound exist for any subset $K \subseteq L$. If $K = L$ they are called 0 (the smallest element in the lattice) and 1 (the biggest element in the lattice) respectively. An *atom* in a lattice is an element a which majorises only 0, i.e., $0 \leq a$ and if $0 \leq b \leq a$ then $b = 0$ or $b = a$. The lattice is called *atomic* if each of its elements different from 0 majorises an atom. An atomic lattice is called *atomistic* if every element is the join of the atoms it majorises. An element c is said to *cover* a if $a < c$ and if $a \leq b \leq c$ either $a = b$ or $b = c$. An atomic lattice is said to have the *covering property* if, for every element b and every atom a for which $a \wedge b = 0$, the join $a \vee b$ covers b.

The subset $\{a, b, c\} \subseteq L$ is called a *distributive triple* if

$$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c) \quad \text{and } (a, b, c) \text{ cyclically permuted}\,, \tag{4.63a}$$

$$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c) \quad \text{and } (a, b, c) \text{ cyclically permuted}\,. \tag{4.63b}$$

Definition 4. A lattice is called *distributive* or *Boolean* if every triple $\{a, b, c\}$ is distributive. It is called *modular* if every triple $\{a, b, c\}$ with $a \leq b$ is distributive.

It is straightforward to check from (4.63) that modularity is equivalent to the following single condition:

$$\text{modularity} \Leftrightarrow a \vee (b \wedge c) = b \wedge (a \vee c) \quad \text{for all } a, b, c \in L \text{ s.t. } a \leq b\,. \tag{4.64}$$

If in a lattice with smallest element 0 and greatest element 1 a map $L \to L$, $a \mapsto a'$, exist such that

$$a'' := (a')' = a\,, \tag{4.65a}$$

$$a \leq b \Rightarrow b' \leq a'\,, \tag{4.65b}$$

$$a \wedge a' = 0\,, \quad a \vee a' = 1\,, \tag{4.65c}$$

[16] By 'equator' we mean the $(n-2)$ – sphere in which the forward and backward light-cones in (4.59) intersect. In the two-dimensional drawings the "equator" is represented by just two points marking the right and left corners of the diamond-shaped set.

the lattice is called *orthocomplemented*. It follows that whenever the meet and join of a subset $\{a_i \mid i \in I\}$ (I is some index set) exist one has De Morgan's laws[17]:

$$\left(\bigwedge_{i\in I} a_i\right)' = \bigvee_{i\in I} a_i' \,, \tag{4.66a}$$

$$\left(\bigvee_{i\in I} a_i\right)' = \bigwedge_{i\in I} a_i' \,. \tag{4.66b}$$

For orthocomplemented lattices there is a still weaker version of distributivity than modularity, which turns out to be physically relevant in various contexts:

Definition 5. An orthocomplemented lattice is called *orthomodular* if every triple $\{a, b, c\}$ with $a \leq b$ and $c \leq b'$ is distributive.

From (4.64) and using that $b \wedge c = 0$ for $b \leq c'$ one sees that this is equivalent to the single condition (renaming c to c'):

$$\text{orthomod.} \Leftrightarrow \quad a = b \wedge (a \vee c') \quad \text{for all } a, b, c \in L \text{ s.t. } a \leq b \leq c \,, \tag{4.67a}$$

$$\Leftrightarrow \quad a = b \vee (a \wedge c') \quad \text{for all } a, b, c \in L \text{ s.t. } a \geq b \geq c \,, \tag{4.67b}$$

where the second line follows from the first by taking its orthocomplement and renaming a', b', c to a, b, c'. It turns out that these conditions can still be simplified by making them independent of c. In fact, (4.67) are equivalent to

$$\text{orthomod.} \Leftrightarrow \quad a = b \wedge (a \vee b') \quad \text{for all } a, b \in L \text{ s.t. } a \leq b \,, \tag{4.68a}$$

$$\Leftrightarrow \quad a = b \vee (a \wedge b') \quad \text{for all } a, b \in L \text{ s.t. } a \geq b \,. \tag{4.68b}$$

It is obvious that (4.67) implies (4.68) (set $c = b$). But the converse is also true. To see this, take e.g., (4.68b) and choose any $c \leq b$. Then $c' \geq b'$, $a \geq b$ (by hypothesis), and $a \geq a \wedge c'$ (trivially), so that $a \geq b \vee (a \wedge c')$. Hence $a \geq b \vee (a \wedge c') \geq b \vee (a \wedge b') = a$, which proves (4.67b).

Complete orthomodular atomic lattices are automatically atomistic. Indeed, let b be the join of all atoms majorised by $a \neq 0$. Assume $a \neq b$ so that necessarily $b < a$, then (4.68b) implies $a \wedge b' \neq 0$. Then there exists an atom c majorised by $a \wedge b'$. This implies $c \leq a$ and $c \leq b'$, hence also $c \not\leq b$. But this is a contradiction, since b is by definition the join of all atoms majorised by a.

Finally we mention the notion of *compatibility* or *commutativity*, which is a symmetric, reflexive, but generally not transitive relation R on an orthomodular lattice (cf. Appendix 1). We write $a \natural b$ for $(a, b) \in R$ and define:

$$a \natural b \quad \Leftrightarrow \quad a = (a \wedge b) \vee (a \wedge b') \,, \tag{4.69a}$$

$$\Leftrightarrow \quad b = (b \wedge a) \vee (b \wedge a') \,. \tag{4.69b}$$

[17] From these laws it also appears that the definition (4.65c) is redundant, as each of its two statements follows from the other, due to $0' = 1$.

The equivalence of these two lines, which shows that the relation of being compatible is indeed symmetric, can be demonstrated using orthomodularity as follows: Suppose (4.69a) holds; then $b \wedge a' = b \wedge (b' \vee a') \wedge (b \vee a') = b \wedge (b' \vee a')$, where we used the orthocomplement of (4.69a) to replace a' in the first expression and the trivial identity $b \wedge (b \vee a') = b$ in the second step. Now, applying (4.68b) to $b \geq a \wedge b$ we get $b = (b \wedge a) \vee [b \wedge (b' \vee a')] = (b \wedge a) \vee (b \wedge a')$, i.e. (4.69b). The converse, (4.69b)$\Rightarrow$(4.69a), is of course entirely analogous.

From (4.69) a few things are immediate: $a \natural b$ is equivalent to $a \natural b'$, $a \natural b$ is implied by $a \leq b$ or $a \leq b'$, and the elements 0 and 1 are compatible with all elements in the lattice. The *centre* of a lattice is the set of elements which are compatible with all elements in the lattice. In fact, the centre is a Boolean sublattice. If the centre contains no other elements than 0 and 1 the lattice is said to be *irreducible*. The other extreme is a Boolean lattice, which is identical to its own centre. Indeed, if (a, b, b') is a distributive triple, one has $a = a \wedge 1 = a \wedge (b \vee b') = (a \wedge b) \vee (a \wedge b') \Rightarrow$(4.69a).

After these digression into elementary notions of lattice theory we come back to our examples of the sets $\mathrm{Caus}(\mathbb{M}^n)$ $\mathrm{Chron}(\mathbb{M}^n)$. Our statements above amount to saying that they are complete, atomic, and orthocomplemented lattices. The partial order relation $\leq$ is given by $\subseteq$ and the extreme elements 0 and 1 correspond to the empty set $\emptyset$ and the total set $\mathbb{M}^n$, the points of which are the atoms. Neither the covering property nor modularity is shared by any of the two lattices, as can be checked by way of elementary counterexamples.[18] In particular, neither of them is Boolean. However, in [15] it was shown that $\mathrm{Chron}(\mathbb{M}^n)$ is orthomodular; see also [13] which deals with more general spacetimes. Note that by the argument given above this implies that $\mathrm{Chron}(\mathbb{M}^n)$ is atomistic. In contrast, $\mathrm{Caus}(\mathbb{M}^n)$ is definitely *not* orthomodular, as is e.g. seen by the counterexample given in Fig. 4.2.[19] It is also not difficult to prove that $\mathrm{Chron}(\mathbb{M}^n)$ is irreducible.[20]

It is well known that the lattices of propositions for classical systems are Boolean, whereas those for quantum systems are merely orthomodular. In classical physics the elements of the lattice are measurable subsets of phase space, with $\leq$ being ordinary set-theoretic inclusion $\subseteq$, and $\wedge$ and $\vee$ being ordinary set-theoretic

[18] An immediate counterexample for the covering property is this: Take two timelike separated points (i.e., atoms) p and q. Then $\{p\} \wedge \{q\} = \emptyset$ whereas $\{p\} \vee \{q\}$ is given by the closed diamond (4.59). Note that this is true in $\mathrm{Caus}(\mathbb{M}^n)$ *and* $\mathrm{Chron}(\mathbb{M}^n)$. But, clearly, $\{p\} \vee \{q\}$ does not cover either $\{p\}$ or $\{q\}$.

[19] Regarding this point, there are some conflicting statements in the literature. The first edition of [27] states orthomodularity of $\mathrm{Chron}(\mathbb{M}^n)$ in Proposition 4.1.3, which is removed in the second edition without further comment. The proof offered in the first edition uses (4.68a) as definition of orthomodularity, writing K_1 for a and K_2 for b. The crucial step is the claim that any spacetime event in the set $K_2 \wedge (K_1 \vee K_2')$ lies in K_2 and that any causal line through it must intersect either K_1 or K_2'. The last statement is, however, not correct since the join of two sets (here K_1 and K_2') is generally larger than the domain of dependence of their ordinary set-theoretic union; compare Fig. 4.2. : (Generally, the domain of dependence of a subset S of spacetime M is the largest subset $D(S) \subseteq M$ such that any inextensible causal curve that intersects $D(S)$ also intersects S.)

[20] In general spacetimes M, the failure of irreducibility of $\mathrm{Chron}(M)$ is directly related to the existence of closed timelike curves; see [13].

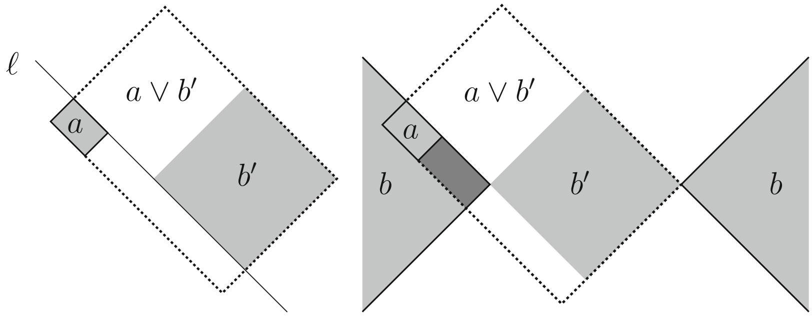


Fig. 4.2 The two figures show that Caus($\mathbb{M}^n$) is not orthomodular. The first thing to note is that Caus($\mathbb{M}^n$) contains open (4.15) as well as closed (4.59) diamond sets. In the left picture we consider the join of a small closed diamond a with a large open diamond b'. (Closed sets are indicated by a solid boundary line.) Their edges are aligned along the lightlike line ℓ. Even though these regions are causally disjoint, their causal completion is much larger than their union and given by the open (for $n > 2$) enveloping diamond $a \vee b'$ framed by the dashed line. (This also shows that the join of two regions can be larger than the domain of dependence of their union; compare footnote 19.). Next we consider the situation depicted on the right side. The closed double-wedge region b contains the small closed diamond a. The causal complement b' of b is the open diamond in the middle. $a \vee b'$ is, according to the first picture, given by the large open diamond enclosed by the dashed line. The intersection of $a \vee b'$ with b is strictly larger than a, the difference being the dark-shaded region in the left wedge of b below a. Hence $a \neq b \wedge (a \vee b')$, in contradiction to (4.68a)

intersection $\cap$ and union $\cup$ respectively. The orthocomplement is the ordinary set-theoretic complement. In Quantum Mechanics the elements of the lattice are the closed subspaces of Hilbert space, with $\leq$ being again ordinary inclusion, $\wedge$ ordinary intersection, and $\vee$ is given by $a \vee b := \overline{\text{span}\{a, b\}}$. The orthocomplement of a closed subset is the orthogonal complement in Hilbert space. For comprehensive discussions see [33] and [4].

One of the main questions in the foundations of Quantum Mechanics is whether one could understand (derive) the usage of Hilbert spaces and complex numbers from somehow more fundamental principles. Even though it is not a priori clear what ones measure of fundamentality should be at this point, an interesting line of attack consists in deriving the mentioned structures from the properties of the lattice of propositions (Quantum Logic). It can be shown that a lattice that is complete, atomic, irreducible, orthomodular, and that satisfies the covering property is isomorphic to the lattice of closed subspaces of a linear space with Hermitean inner product. The complex numbers are selected if additional technical assumptions are added. For the precise statements of these reconstruction theorems see [4].

It is now interesting to note that, on a formal level, there is a similar transition in going from Galilei invariant to Lorentz invariant causality relations. In fact, in Galilean spacetime one can also define a chronological complement: Two points are chronologically related if they are connected by a worldline of finite speed and, accordingly, two subsets in spacetime are chronologically disjoint if no point in one set is chronologically related to a point of the other. For example, the chronological complement of a point p are all points simultaneous to, but different from, p. More general, it is not hard to see that the chronologically complete sets are just the subsets of some $t = \text{const.}$ hypersurface. The lattice of chronologically complete sets is then the continuous disjoint union of sublattices, each of which is isomorphic to the Boolean lattice of subsets in $\mathbb{R}^3$. For details see [14].

As we have seen above, Chron($\mathbb{M}^n$) is complete, atomic, irreducible, and orthomodular (hence atomistic). The main difference to the lattice of propositions in Quantum Mechanics, as regards the formal aspects discussed here, is that Chron($\mathbb{M}^n$) does *not* satisfy the covering property. Otherwise the formal similarities are intriguing and it is tempting to ask whether there is a deeper meaning to this. In this respect it would be interesting to know whether one could give a lattice-theoretic characterisation for Chron(M) (M some fixed spacetime), comparable to the characterisation of the lattices of closed subspaces in Hilbert space alluded to above. Even for $M = \mathbb{M}^n$ such a characterisation seems, as far as I am aware, not to be known.

4.3.3 Rigid Motion

As is well known, the notion of a rigid body, which proves so useful in Newtonian mechanics, is incompatible with the existence of a universal finite upper bound for all signal velocities [36]. As a result, the notion of a perfectly rigid body does not exist within the framework of SR. However, the notion of a *rigid motion* does exist. Intuitively speaking, a body moves rigidly if, locally, the relative spatial distances of its material constituents are unchanging.

The motion of an extended body is described by a normalised timelike vector field $u : \Omega \to \mathbb{R}^n$, where Ω is an open subset of Minkowski space, consisting of the events where the material body in question "exists". We write $g(u, u) = u \cdot u = u^2$ for the Minkowskian scalar product. Being normalised now means that $u^2 = c^2$ (we do *not* choose units such that $c = 1$). The Lie derivative with respect to u is denoted by L_u.

For each material part of the body in motion its local rest space at the event $p \in \Omega$ can be identified with the hyperplane through p orthogonal to u_p:

$$H_p := p + u_p^\perp \,. \tag{4.70}$$

$u_p^\perp$ carries a Euclidean inner product, h_p, given by the restriction of $-g$ to $u_p^\perp$. Generally we can write

$$h = c^{-2}\, u^\flat \otimes u^\flat - g \,, \tag{4.71}$$

where $u^\flat = g^\downarrow(u) := g(u, \cdot)$ is the one-form associated to u. Following [9] the precise definition of "rigid motion" can now be given as follows:

Definition 6 (Born 1909). Let u be a normalised timelike vector field u. The motion described by its flow is *rigid* if

$$L_u h = 0 \,. \tag{4.72}$$

Note that, in contrast to the Killing equations $L_u g = 0$, these equations are non linear due to the dependence of h upon u.

We write $\Pi_h := \mathrm{id} - c^{-2}\, u \otimes u^\flat \in \mathrm{End}(\mathbb{R}^n)$ for the tensor field over spacetime that pointwise projects vectors perpendicular to u. It acts on one forms α via $\Pi_h(\alpha) := \alpha \circ \Pi_h$ and accordingly on all tensors. The so extended projection map will still be denoted by Π_h. Then we, e.g., have

$$h = -\Pi_h g := -g(\Pi_h \cdot, \Pi_h \cdot)\,. \tag{4.73}$$

It is not difficult to derive the following two equations:[21]

$$L_{fu} h = f L_u h\,, \tag{4.74}$$
$$L_u h = -L_u(\Pi_h g) = -\Pi_h(L_u g)\,, \tag{4.75}$$

where f is any differentiable real-valued function on Ω.

Equation (4.74) shows that the normalised vector field u satisfies (4.72) iff any rescaling fu with a nowhere vanishing function f does. Hence the normalization condition for u in (4.72) is really irrelevant. It is the geometry in spacetime of the flow lines and not their parameterisation which decide on whether motions (all, i.e., for any parameterisation, or none) along them are rigid. This has be the case because, generally speaking, there is no distinguished family of sections (hypersurfaces) across the bundle of flow lines that would represent "the body in space", i.e. mutually simultaneous locations of the body's points. Distinguished cases are those exceptional ones in which u is hypersurface orthogonal. Then the intersection of u's flow lines with the orthogonal hypersurfaces consist of mutually *Einstein synchronous* locations of the points of the body. An example is discussed below.

Equation (4.75) shows that the rigidity condition is equivalent to the "spatially" projected Killing equation. We call the flow of the timelike normalised vector field u a *Killing motion* (i.e., a spacetime isometry) if there is a Killing field K such that $u = cK/\sqrt{K^2}$. Equation (4.75) immediately implies that Killing motions are rigid. What about the converse? Are there rigid motions that are not Killing? This turns out to be a difficult question. Its answer in Minkowski space is: "yes, many, but not as many as naïvely expected."

Before we explain this, let us give an illustrative example for a Killing motion, namely that generated by the boost Killing-field in Minkowski space. We suppress all but one spatial directions and consider boosts in x direction in two-dimensional

[21] Equation (4.75) simply follows from $L_u \Pi_h = -c^{-2} u \otimes L_u u^\flat$, so that $g((L_u\Pi_h)X, \Pi_h Y) = 0$ for all X, Y. In fact, $L_u u^\flat = a^\flat$, where $a := \nabla_u u$ is the spacetime-acceleration. This follows from $L_u u^\flat(X) = L_u(g(u,X)) - g(u, L_u X) = g(\nabla_u u, X) + g(u, \nabla_u X - [u,X]) = g(a,X) - g(u, \nabla_X u) = g(a,X)$, where $g(u,u) = \text{const.}$ was used in the last step.

Minkowski space (coordinates ct and x; metric $ds^2 = c^2dt^2 - dx^2$). The Killing field is[22]

$$K = x\,\partial_{ct} + ct\,\partial_x\,, \tag{4.76}$$

which is timelike in the region $|x| > |ct|$. We focus on the "right wedge" $x > |ct|$, which is now our region Ω. Consider a rod of length ℓ which at $t = 0$ is represented by the interval $x \in (r, r+\ell)$, where $r > 0$. The flow of the normalised field $u = cK/\sqrt{K^2}$ is

$$ct(\tau) \;=\; x_0\,\sinh\bigl(c\tau/x_0\bigr)\,, \tag{4.77a}$$

$$x(\tau) \;=\; x_0\,\cosh\bigl(c\tau/x_0\bigr)\,, \tag{4.77b}$$

where $x_0 = x(\tau = 0) \in (r, r+\ell)$ labels the elements of the rod at $\tau = 0$. We have $x^2 - c^2t^2 = x_0^2$, showing that the individual elements of the rod move on hyperbolae ("hyperbolic motion"). τ is the proper time along each orbit, normalised so that the rod lies on the x axis at $\tau = 0$.

The combination

$$\lambda := c\tau/x_0 \tag{4.78}$$

is just the flow parameter for K (4.76), sometimes referred to as "Killing time" (though it is dimensionless). From (4.77) we can solve for λ and τ as functions of ct and x:

$$\lambda \;=\; f(ct, x) \;:=\; \tanh^{-1}\bigl(ct/x\bigr)\,, \tag{4.79a}$$

$$\tau \;=\; \hat{f}(ct, x) \;:=\; \underbrace{\sqrt{(x/c)^2 - t^2}}_{x_0/c}\,\tanh^{-1}\bigl(ct/x\bigr)\,, \tag{4.79b}$$

from which we infer that the hypersurfaces of constant λ are hyperplanes which all intersect at the origin. Moreover, we also have $df = K^\flat/K^2$ (d is just the ordinary exterior differential) so that the hyperplanes of constant λ intersect all orbits of u (and K) orthogonally. Hence the hyperplanes of constant λ qualify as the equivalence classes of mutually Einstein-simultaneous events in the region $x > |ct|$ for a family of observers moving along the Killing orbits. This does not hold for the hypersurfaces of constant τ, which are curved.

The modulus of the spacetime-acceleration (which is the same as the modulus of the spatial acceleration measured in the local rest frame) of the material part of the rod labelled by x_0 is

$$\|a\|_g = c^2/x_0\,. \tag{4.80}$$

As an aside we generally infer from this that, given a timelike curve of local acceleration (modulus) α, infinitesimally nearby orthogonal hyperplanes intersect at a

[22] Here we adopt the standard notation from differential geometry, where $\partial_\mu := \partial/\partial x^\mu$ denote the vector fields naturally defined by the coordinates $\{x^\mu\}_{\mu=0\cdots n-1}$. Pointwise the dual basis to $\{\partial_\mu\}_{\mu=0\cdots n-1}$ is $\{dx^\mu\}_{\mu=0\cdots n-1}$.

spatial distance c^2/α. This remark will become relevant in the discussion of part 2 of the Noether-Herglotz theorem given below.

In order to accelerate the rod to the uniform velocity v without deforming it, its material point labelled by x_0 has to accelerate for the eigentime (this follows from (4.77))

$$\tau = \frac{x_0}{c} \tanh^{-1}(v/c)\,, \tag{4.81}$$

which depends on x_0. In contrast, the Killing time is the same for all material points and just given by the final rapidity. In particular, judged from the local observers moving with the rod, a rigid acceleration requires accelerating the rod's trailing end harder but shorter than pulling its leading end.

In terms of the coordinates (λ, x_0), which are co-moving with the flow of K, and (τ, x_0), which are co-moving with the flow of u, we just have $K = \partial/\partial\lambda$ and $u = \partial/\partial\tau$ respectively. The spacetime metric g and the projected metric h in terms of these coordinates are:

$$h \;=\; dx_0^2\,, \tag{4.82a}$$

$$g \;=\; x_0^2\, d\lambda^2 - dx_0^2 = c^2\big(d\tau - (\tau/x_0)\, dx_0\big)^2 - dx_0^2\,. \tag{4.82b}$$

Note the simple form g takes in terms of x_0 and λ, which are also called the 'Rindler coordinates' for the region $|x| > |ct|$ of Minkowski space. They are the analogs in Lorentzian geometry to polar coordinates (radius x_0, angle λ) in Euclidean geometry.

Let us now return to the general case. We decompose the derivative of the velocity one-form $u^\flat := g^\downarrow(u)$ as follows:

$$\nabla u^\flat = \theta + \omega + c^{-2}\, u^\flat \otimes a^\flat\,, \tag{4.83}$$

where θ and ω are the projected symmetrised and antisymmetrised derivatives respectively[23]

$$2\theta \;=\; \Pi_h(\nabla \vee u^\flat) \;=\; \nabla \vee u^\flat - c^{-2}\, u^\flat \vee a^\flat\,, \tag{4.84a}$$

$$2\omega \;=\; \Pi_h(\nabla \wedge u^\flat) \;=\; \nabla \wedge u^\flat - c^{-2}\, u^\flat \wedge a^\flat\,. \tag{4.84b}$$

The symmetric part, θ, is usually further decomposed into its traceless and pure trace part, called the *shear* and *expansion* of u respectively. The antisymmetric part ω is called the *vorticity* of u.

[23] We denote the symmetrised and antisymmetrised tensor-product (not including the factor $1/n!$) by $\vee$ and $\wedge$ respectively and the symmetrised and antisymmetrised (covariant-) derivative by $\nabla\vee$ and $\nabla\wedge$. For example, $(u^\flat \wedge v^\flat)_{ab} = u_a v_b - u_b v_a$ and $(\nabla \vee u^\flat)_{ab} = \nabla_a u_b + \nabla_b u_a$. Note that $(\nabla \wedge u^\flat)$ is the same as the ordinary exterior differential $du^\flat$. Everything we say in the sequel applies to curved spacetimes if ∇ is read as covariant derivative with respect to the Levi-Civita connection.

Now recall that the Lie derivative of g is just twice the symmetrised derivative, which in our notation reads:

$$L_u g = \nabla \vee u^\flat \, . \tag{4.85}$$

This implies in view of (4.72), (4.75), and (4.84a)

Proposition 6. *Let u be a normalised timelike vector field u. The motion described by its flow is rigid iff u is of vanishing shear and expansion, i.e., iff* $\theta = 0$.

Vector fields generating rigid motions are now classified according to whether or not they have a vanishing vorticity ω: if $\omega = 0$ the flow is called *irrotational*, otherwise *rotational*. The following theorem is due to Herglotz [29] and Noether [40]:

Theorem 8 (Noether & Herglotz, part 1). *A rotational rigid motion in Minkowski space must be a Killing motion.*

An example of such a rotational motion is given by the Killing field[24]

$$K = \partial_t + \kappa \, \partial_\varphi \tag{4.86}$$

inside the region

$$\Omega = \{(t, z, \rho, \varphi) \mid \kappa\rho < c\} \, , \tag{4.87}$$

where K is timelike. This motion corresponds to a rigid rotation with constant angular velocity κ which, without loss of generality, we take to be positive. Using the co-moving angular coordinate $\psi := \varphi - \kappa t$, the split (4.71) is now furnished by

$$u^\flat = c\,\sqrt{1 - (\kappa\rho/c)^2}\left\{c\,dt - \frac{\kappa\rho/c}{1 - (\kappa\rho/c)^2}\,\rho\,d\psi\right\} \, , \tag{4.88a}$$

$$h = dz^2 + d\rho^2 + \frac{\rho^2\,d\psi^2}{1 - (\kappa\rho/c)^2} \, . \tag{4.88b}$$

The metric h is curved (cf. Lemma 4). But the rigidity condition (4.72) means that h, and hence its curvature, cannot change along the motion. Therefore, even though we can keep a body in uniform rigid rotational motion, we cannot put it into this state from rest by purely rigid motions, since this would imply a transition from a flat to a curved geometry of the body. This was first pointed out by Ehrenfest [19]. Below we will give a concise analytical expression of this fact (cf. Eq. (4.92)). All this is in contrast to the translational motion, as we will also see below.

The proof of Theorem 8 relies on arguments from differential geometry proper and is somewhat tricky. Here we present the essential steps, basically following [42] and [46] in a slightly modernised notation. Some straightforward calculational details will be skipped. The argument itself is best broken down into several lemmas.

---

[24] We now use standard cylindrical coordinates (z, ρ, φ), in terms of which $ds^2 = c^2 dt^2 - dz^2 - d\rho^2 - \rho^2\,d\varphi^2$.

At the heart of the proof lies the following general construction: Let M be the spacetime manifold with metric g and $\Omega \subset M$ the open region in which the normalised vector field u is defined. We take Ω to be simply connected. The orbits of u foliate Ω and hence define an equivalence relation on Ω given by $p \sim q$ iff p and q lie on the same orbit. The quotient space $\hat{\Omega} := \Omega/\!\sim$ is itself a manifold. Tensor fields on $\hat{\Omega}$ can be represented by (i.e., are in bijective correspondence to) tensor fields T on Ω which obey the two conditions:

$$\Pi_h T = T\,, \tag{4.89a}$$
$$L_u T = 0\,. \tag{4.89b}$$

Tensor fields satisfying (4.89a) are called *horizontal*, those satisfying both conditions (4.89) are called *projectable*. The $(n-1)$-dimensional metric tensor h, defined in (4.71), is an example of a projectable tensor if u generates a rigid motion, as assumed here. It turns $(\hat{\Omega}, h)$ into a $(n-1)$-dimensional Riemannian manifold. The covariant derivative $\hat{\nabla}$ with respect to the Levi-Civita connection of h is given by the following operation on projectable tensor fields:

$$\hat{\nabla} := \Pi_h \circ \nabla \tag{4.90}$$

i.e., by first taking the covariant derivative ∇ (Levi-Civita connection in (M, g)) in spacetime and then projecting the result horizontally. This results again in a projectable tensor, as a straightforward calculation shows.

The horizontal projection of the spacetime curvature tensor can now be related to the curvature tensor of $\hat{\Omega}$ (which is a projectable tensor field). Without proof we state

Lemma 4. *Let u generate a rigid motion in spacetime. Then the horizontal projection of the totally covariant (i.e., all indices down) curvature tensor R of (Ω, g) is related to the totally covariant curvature tensor $\hat{R}$ of $(\hat{\Omega}, h)$ by the following equation[25]:*

$$\Pi_h R = -\hat{R} - 3\,(\mathrm{id} - \Pi_\wedge)\omega \otimes \omega\,, \tag{4.91}$$

where $\Pi_\wedge$ is the total antisymmetriser, which here projects tensors of rank four onto their totally antisymmetric part.

Formula (4.91) is true in any spacetime dimension n. Note that the projector $(\mathrm{id} - \Pi_\wedge)$ guarantees consistency with the first Bianchi identities for R and $\hat{R}$, which state that the total antisymmetrisation in their last three slots vanish identically. This is consistent with (4.91) since for tensors of rank four with the symmetries of $\omega \otimes \omega$ the total antisymmetrisation on tree slots is identical to $\Pi_\wedge$, the symmetrisation on all four slots. The claim now simply follows from $\Pi_\wedge \circ (\mathrm{id} - \Pi_\wedge) = \Pi_\wedge - \Pi_\wedge = 0$.

---

[25] $\hat{R}$ appears with a minus sign on the right hand side of (4.91) because the first index on the hatted curvature tensor is lowered with h rather than g. This induces a minus sign due to (4.71), i.e., as a result of our "mostly-minus"-convention for the signature of the spacetime metric.

We now restrict to spacetime dimensions of four or less, i.e. $n \leq 4$. In this case $\Pi_\wedge \circ \Pi_h = 0$ since Π_h makes the tensor effectively live over $n-1$ dimensions, and any totally antisymmetric four-tensor in three or less dimensions must vanish. Applied to (4.91) this means that $\Pi_\wedge(\omega \otimes \omega) = 0$, for horizontality of ω implies $\omega \otimes \omega = \Pi_h(\omega \otimes \omega)$. Hence the right hand side of (4.91) just contains the pure tensor product $-3\,\omega \otimes \omega$.

Now, in our case $R = 0$ since (M, g) is flat Minkowski space. This has two interesting consequences: First, $(\hat{\Omega}, h)$ is curved iff the motion is rotational, as exemplified above. Second, since $\hat{R}$ is projectable, its Lie derivative with respect to u vanishes. Hence (4.91) implies $L_u\omega \otimes \omega + \omega \otimes L_u\omega = 0$, which is equivalent to[26]

$$L_u\omega = 0\,. \tag{4.92}$$

This says that the vorticity cannot change along a rigid motion in flat space. It is the precise expression for the remark above that you cannot rigidly *set* a disk into rotation. Note that it also provides the justification for the global classification of rigid motions into rotational and irrotational ones.

A sharp and useful criterion for whether a rigid motion is Killing or not is given by the following

Lemma 5. *Let u be a normalised timelike vector field on a region $\Omega \subseteq M$. The motion generated by u is Killing iff it is rigid and $a^\flat$ is exact on Ω.*

Proof. That the motion generated by u be Killing is equivalent to the existence of a positive function $f : \Omega \to \mathbb{R}$ such that $L_{fu}g = 0$, i.e. $\nabla \vee (f u^\flat) = 0$. In view of (4.84a) this is equivalent to

$$2\theta + (d \ln f + c^{-2} a^\flat) \vee u^\flat = 0\,, \tag{4.93}$$

which, in turn, is equivalent to $\theta = 0$ and $a^\flat = -c^2\, d \ln f$. This is true since θ is horizontal, $\Pi_h\theta = \theta$, whereas the first term in (4.93) vanishes upon applying Π_h. The result now follows from reading this equivalence both ways: (1) The Killing condition for $K := f u$ implies rigidity for u and exactness of $a^\flat$. (2) Rigidity of u and $a^\flat = -d\Phi$ imply that $K := f u$ is Killing, where $f := \exp(\Phi/c^2)$.

We now return to the condition (4.92) and express $L_u\omega$ in terms of $du^\flat$. For this we recall that $L_u u^\flat = a^\flat$ (cf. footnote 21) and that Lie derivatives on forms commute with exterior derivatives.[27] Hence we have

$$2\, L_u\omega = L_u(\Pi_h du^\flat) = \Pi_h da^\flat = da^\flat - c^{-2} u^\flat \wedge L_u a^\flat\,. \tag{4.94}$$

[26] In more than four spacetime dimensions one only gets $(\mathrm{id} - \Pi_\wedge)(L_u\omega \otimes \omega + \omega \otimes L_u\omega) = 0$.

[27] This is most easily seen by recalling that on forms the Lie derivative can be written as $L_u = d \circ i_u + i_u \circ d$, where i_u is the map of inserting u in the first slot.

Here we used the fact that the additional terms that result from the Lie derivative of the projection tensor Π_h vanish, as a short calculation shows, and also that on forms the projection tensor Π_h can be written as $\Pi_h = \mathrm{id} - c^{-2}u^\flat \wedge i_u$, where i_u denotes the map of insertion of u in the first slot.

Now we prove

Lemma 6. *Let u generate a rigid motion in flat space such that $\omega \neq 0$, then*

$$L_u a^\flat = 0\,. \tag{4.95}$$

Proof. Equation (4.92) says that ω is projectable (it is horizontal by definition). Hence $\hat{\nabla}\omega$ is projectable, which implies

$$L_u \hat{\nabla}\omega = 0\,. \tag{4.96}$$

Using (4.83) with $\theta = 0$ one has

$$\hat{\nabla}\omega = \Pi_h \nabla\omega = \Pi_h \nabla\nabla u^\flat - c^{-2}\Pi_h(\nabla u^\flat \otimes a^\flat)\,. \tag{4.97}$$

Antisymmetrisation in the first two tensor slots makes the first term on the right vanish due to the flatness on ∇. The antisymmetrised right hand side is hence equal to $-c^{-2}\omega \otimes a^\flat$. Taking the Lie derivative of both sides makes the left hand side vanish due to (4.96), so that

$$L_u(\omega \otimes a^\flat) = \omega \otimes L_u a^\flat = 0 \tag{4.98}$$

where we also used (4.92). So we see that $L_u a^\flat = 0$ if $\omega \neq 0$.[28]

The last three lemmas now constitute a proof for Theorem 8. Indeed, using (4.95) in (4.94) together with (4.92) shows $da^\flat = 0$, which, according to Lemma 5, implies that the motion is Killing.

Next we turn to the second part of the theorem of Noether and Herglotz, which reads as follows:

Theorem 9 (Noether & Herglotz, part 2). *All irrotational rigid motions in Minkowski space are given by the following construction: take a twice continuously differentiable curve $\tau \mapsto z(\tau)$ in Minkowski space, where w.l.o.g τ is the eigentime, so that $\dot{z}^2 = c^2$. Let $H_\tau := z(\tau) + (\dot{z}(\tau))^\perp$ be the hyperplane through $z(\tau)$ intersecting the curve z perpendicularly. Let Ω be a the tubular neighbourhood of z in which no two hyperplanes $H_\tau, H_{\tau'}$ intersect for any pair $z(\tau), z(\tau')$ of points on the curve. In Ω define u as the unique (once differentiable) normalised timelike vector field perpendicular to all $H_\tau \cap \Omega$. The flow of u is the sought-for rigid motion.*

[28] We will see below that (4.95) is generally not true if $\omega = 0$; see Eq. (4.107).

Proof. We first show that the flow so defined is indeed rigid, even though this is more or less obvious from its very definition, since we just defined it by "rigidly" moving a hyperplane through spacetime. In any case, analytically we have,

$$H_\tau = \{x \in \mathbb{M}^n \mid f(\tau, x) := \dot{z}(\tau) \cdot \big(x - z(\tau)\big) = 0\}\,. \tag{4.99}$$

In Ω any x lies on exactly one such hyperplane, H_τ, which means that there is a function $\sigma : \Omega \to \mathbb{R}$ so that $\tau = \sigma(x)$ and hence $F(x) := f(\sigma(x), x) \equiv 0$. This implies $dF = 0$. Using the expression for f from (4.99) this is equivalent to

$$d\sigma = \dot{z}^\flat \circ \sigma / [c^2 - (\ddot{z} \circ \sigma) \cdot (\mathrm{id} - z \circ \sigma)]\,, \tag{4.100}$$

where "id" denotes the "identity vector-field", $x \mapsto x^\mu \partial_\mu$, in Minkowski space. Note that in Ω we certainly have $\partial_\tau f(\tau, x) \neq 0$ and hence $\ddot{z} \cdot (x - z) \neq c^2$. In Ω we now define the normalised timelike vector field[29]

$$u := \dot{z} \circ \sigma\,. \tag{4.101}$$

Using (4.100), its derivative is given by

$$\nabla u^\flat = d\sigma \otimes (\ddot{z}^\flat \circ \sigma) = \big[(\dot{z}^\flat \circ \sigma) \otimes (\ddot{z}^\flat \circ \sigma)\big]/(N^2 c^2)\,, \tag{4.102}$$

where

$$N := 1 - (\ddot{z} \circ \sigma) \cdot (\mathrm{id} - z \circ \sigma)/c^2\,. \tag{4.103}$$

This immediately shows that $\Pi_h \nabla u^\flat = 0$ (since $\Pi_h \dot{z}^\flat = 0$) and therefore that $\theta = \omega = 0$. Hence u, as defined in (4.101), generates an irrotational rigid motion.

For the converse we need to prove that any irrotational rigid motion is obtained by such a construction. So suppose u is a normalised timelike vector field such that $\theta = \omega = 0$. Vanishing ω means $\Pi_h(\nabla \wedge u^\flat) = \Pi_h(du^\flat) = 0$. This is equivalent to $u^\flat \wedge du^\flat = 0$, which according to the Frobenius theorem in differential geometry is equivalent to the integrability of the distribution[30] $u^\flat = 0$, i.e., the hypersurface orthogonality of u. We wish to show that the hypersurfaces orthogonal to u are hyperplanes. To this end consider a spacelike curve $z(s)$, where s is the proper length, running within one hypersurface perpendicular to u. The component of its second s-derivative parallel to the hypersurface is given by (to save notation we now simply write u and $u^\flat$ instead of $u \circ z$ and $u^\flat \circ z$)

$$\Pi_h \ddot{z} = \ddot{z} - c^{-2} u\, u^\flat(\ddot{z}) = \ddot{z} + c^{-2} u\, \theta(\dot{z}, \dot{z}) = \ddot{z}\,, \tag{4.104}$$

---

[29] Note that, by definition of σ, $(\dot{z} \circ \sigma) \cdot (\mathrm{id} - z \circ \sigma) \equiv 0$.

[30] "Distribution" is here used in the differential-geometric sense, where for a manifold M it denotes an assignment of a linear subspace V_p in the tangent space $T_p M$ to each point p of M. The distribution $u^\flat = 0$ is defined by $V_p = \{v \in T_p M \mid u^\flat_p(v) = u_p \cdot v = 0\}$. A distribution is called (locally) integrable if (in the neighbourhood of each point) there is a submanifold M' of M whose tangent space at any $p \in M'$ is just V_p.

where we made a partial differentiation in the second step and then used $\theta = 0$. Geodesics in the hypersurface are curves whose second derivative with respect to proper length have vanishing components parallel to the hypersurface. Now, (4.104) implies that geodesics in the hypersurface are geodesics in Minkowski space (the hypersurface is "totally geodesic"), i.e., given by straight lines. Hence the hypersurfaces are hyperplanes.

Theorem 9 precisely corresponds to the Newtonian counterpart: The irrotational motion of a rigid body is determined by the worldline of any of its points, and any timelike worldline determines such a motion. We can rigidly put an extended body into any state of translational motion, as long as the size of the body is limited by c^2/α, where α is the modulus of its acceleration. This also shows that (4.95) is generally not valid for irrotational rigid motions. In fact, the acceleration one-form field for (4.101) is

$$a^\flat = (\ddot{z}^\flat \circ \sigma)/N \tag{4.105}$$

from which one easily computes

$$da^\flat = (\dot{z}^\flat \circ \sigma) \wedge \left\{ (\Pi_h \dddot{z}^\flat \circ \sigma) + (\ddot{z}^\flat \circ \sigma) \frac{(\Pi_h \ddot{z} \circ \sigma) \cdot (\mathrm{id} - z \circ \sigma)}{Nc^2} \right\} N^{-2} c^{-2} \,. \tag{4.106}$$

From this one sees, for example, that for *constant acceleration*, defined by $\Pi_h \dddot{z} = 0$ (constant acceleration in time as measured in the instantaneous rest frame), we have $da^\flat = 0$ and hence a Killing motion. Clearly, this is just the motion (4.77) for the boost Killing field (4.76). The Lie derivative of $a^\flat$ is now easily obtained:

$$L_u a^\flat = i_u da^\flat = (\Pi_h \dddot{z}^\flat \circ \sigma) N^{-2} \,, \tag{4.107}$$

showing explicitly that it is not zero except for motions of constant acceleration, which were just seen to be Killing motions.

In contrast to the irrotational case just discussed, we have seen that we cannot put a body rigidly into rotational motion. In the old days this was sometimes expressed by saying that the rigid body in SR has only three instead of six degrees of freedom. This was clearly thought to be paradoxical as long as one assumed that the notion of a perfectly rigid body should also make sense in the framework of SR. However, this hope was soon realized to be physically untenable [36].

Appendices

In this appendix we spell out in detail some of the mathematical notions that were used in the main text.

Appendix 1: Sets and Group Actions

Given a set S, recall that an *equivalence relation* is a subset $R \subset S \times S$ such that for all $p, q, r \in S$ the following conditions hold: (1) $(p, p) \in R$ (called "reflexivity"), (2) if $(p, q) \in R$ then $(q, p) \in R$ (called "symmetry"), and (3) if $(p, q) \in R$ and $(q, r) \in R$ then $(p, r) \in R$ (called "transitivity"). Once R is given, one often conveniently writes $p \sim q$ instead of $(p, q) \in R$. Given $p \in S$, its *equivalence class*, $[p] \subseteq S$, is given by all points R-related to p, i.e. $[p] := \{q \in S \mid (p, q) \in R\}$. One easily shows that equivalence classes are either identical or disjoint. Hence they form a *partition* of S, that is, a covering by mutually disjoint subsets. Conversely, given a partition of a set S, it defines an equivalence relation by declaring two points as related iff they are members of the same cover set. Hence there is a bijective correspondence between partitions of and equivalence relations on a set S. The set of equivalence classes is denoted by S/R or $S/\sim$. There is a natural surjection $S \to S/R$, $p \mapsto [p]$.

If in the definition of equivalence relation we exchange symmetry for antisymmetry, i.e., $(p, q) \in R$ and $(q, p) \in R$ implies $p = q$, the relation is called a *partial order*, usually written as $p \geq q$ for $(p, q) \in R$. If, instead, reflexivity is dropped and symmetry is replaced by asymmetry, i.e., $(p, q) \in R$ implies $(q, p) \notin R$, one obtains a relation called a *strict partial order*, usually denoted by $p > q$ for $(p, q) \in R$.

An *left action* of a group G on a set S is a map $\phi : G \times S \to S$, such that $\phi(e, s) = s$ (e = group identity) and $\phi(gh, s) = \phi(g, \phi(h, s))$. If instead of the latter equation we have $\phi(gh, s) = \phi(h, \phi(g, s))$ one speaks of a *right action*. For left actions one sometimes conveniently writes $\phi(g, s) =: g \cdot s$, for right actions $\phi(g, s) =: s \cdot g$. An action is called *transitive* if for every pair $(s, s') \in S \times S$ there is a $g \in G$ such that $\phi(g, s) = s'$, and *simply transitive* if, in addition, (s, s') determine g uniquely, that is, $\phi(g, s) = \phi(g', s)$ for some s implies $g = g'$. The action is called *effective* if $\phi(g, s) = s$ for all s implies $g = e$ ('every $g \neq e$ moves something') and *free* if $\phi(g, s) = s$ for some s implies $g = e$ ('no $g \neq e$ has a fixed point'). It is obvious that simple transitivity implies freeness and that, conversely, freeness and transitivity implies simple transitivity. Moreover, for Abelian groups, effectivity and transitivity suffice to imply simple transitivity. Indeed, suppose $g \cdot s = g' \cdot s$ holds for some $s \in S$, then we also have $k \cdot (g \cdot s) = k \cdot (g' \cdot s)$ for all $k \in G$ and hence $g \cdot (k \cdot s) = g' \cdot (k \cdot s)$ by commutativity. This implies that $g \cdot s = g' \cdot s$ holds, in fact, for all s.

For any $s \in S$ we can consider the *stabilizer subgroup*

$$\mathrm{Stab}(s) := \{g \in G \mid \phi(g, s) = s\} \subseteq G\,. \tag{4.108}$$

If ϕ is transitive, any two stabilizer subgroups are conjugate: $\mathrm{Stab}(g \cdot s) = g\mathrm{Stab}(s)g^{-1}$. By definition, if ϕ is free all stabilizer subgroups are trivial (consist of the identity element only). In general, the intersection $G' := \bigcap_{s \in S} \mathrm{Stab}(s) \subseteq G$ is the normal subgroup of elements acting trivially on S. If ϕ is an action of G on S, then there is an effective action $\hat{\phi}$ of $\hat{G} := G/G'$ on S, defined by $\hat{\phi}([g], s) := \phi(g, s)$, where $[g]$ denotes the G'-coset of G' in G.

The *orbit* of s in S under the action ϕ of G is the subset

$$\mathrm{Orb}(s) := \{\phi(g,s) \mid g \in G\} \subseteq S\,. \tag{4.109}$$

It is easy to see that group orbits are either disjoint or identical. Hence they define a partition of S, that is, an equivalence relation.

A relation R on S is said to be invariant under the self map $f : S \to S$ if $(p,q) \in R \Leftrightarrow (f(p), f(q)) \in R$. It is said to be invariant under the action ϕ of G on S if $(p,q) \in R \Leftrightarrow (\phi(g,p), \phi(g,q)) \in R$ for all $g \in G$. If R is such a G-invariant equivalence relation, there is an action ϕ' of G on the set S/R of equivalence classes, defined by $\phi'(g,[p]) := [\phi(g,p)]$. A general theorem states that invariant equivalence relations exist for transitive group actions, iff the stabilizer subgroups (which in the transitive case are all conjugate) are maximal (e.g. Theorem 1.12 in [31]).

Appendix 2: Affine Spaces

Definition 7. An n-dimensional **affine space** over the field $\mathbb{F}$ (usually $\mathbb{R}$ or $\mathbb{C}$) is a triple (S, V, Φ), where S is a non-empty set, V an n-dimensional vector space over $\mathbb{F}$, and Φ an effective and transitive action $\Phi : V \times S \to S$ of V (considered as Abelian group with respect to addition of vectors) on S.

We remark that an effective and transitive action of an Abelian group is necessarily simply transitive. Hence, without loss of generality, we could have required a simply transitive action in Definition 7 straightaway. We also note that even though the action Φ only refers to the Abelian group structure of V, it is nevertheless important for the definition of an affine space that V is, in fact, a vector space (see below). Any ordered pair of points $(p,q) \in S \times S$ uniquely defines a vector v, namely that for which $p = q + v$. It can be thought of as the difference vector pointing from q to p. We write $v = \Delta(q,p)$, where $\Delta : S \times S \to V$ is a map which satisfies the conditions

$$\Delta(p,q) + \Delta(q,r) = \Delta(p,r) \qquad \text{for all} \quad p,q,r \in S\,, \tag{4.110a}$$

$$\Delta_q : p \ni S \mapsto \Delta(p,q) \in V \quad \text{is a bijection} \qquad \text{for all} \quad p \in S\,. \tag{4.110b}$$

Conversely, these conditions suffice to characterise an affine space, as stated in the following proposition, the proof of which is left to the reader:

Proposition 7. *Let S be a non-empty set, V an n-dimensional vector space over $\mathbb{F}$ and $\Delta : S \times S \to V$ a map satisfying conditions (4.110). Then S is an n-dimensional affine space over $\mathbb{F}$ with action $\Phi(v,p) := \Delta_p^{-1}(v)$.*

One usually writes $\Phi(v,p) =: p + v$, which defines what is meant by "+" between an element of an affine space and an element of V. Note that addition of two points in affine space is not defined. The property of being an action now states $p + 0 = p$ and $(p + v) + w = p + (v + w)$, so that in the latter case we may just write $p + v + w$. Similarly we write $\Delta(p, q) =: q - p$, defining what is meant by "–" between two elements of affine space. The minus sign also makes sense between an element of affine space and an element of vector space if one defines $p + (-v) =: p - v$. We may now write equations like

$$p + (q - r) = q + (p - r)\,, \tag{4.111}$$

the formal proof of which is again left to the reader. It implies that

Considered as Abelian group, any linear subspace $W \subset V$ defines a subgroup. The orbit of that subgroup in S through $p \in S$ is an affine subspace, denoted by W_p, i.e.

$$W_p = p + W := \{p + w \mid w \in W\}\,, \tag{4.112}$$

which is an affine space over W in its own right of dimension $\dim(W)$. One-dimensional affine subspaces are called *(straight) lines*, two-dimensional ones *planes*, and those of co-dimension one are called *hyperplanes*.

Appendix 3: Affine Maps

Affine morphisms, or simply affine maps, are structure preserving maps between affine spaces. To define them in view of Definition 7 we recall once more the significance of V being a vector space and not just an Abelian group. This enters the following definition in an essential way, since there are considerably more automorphisms of V as Abelian group, i.e., maps $f : V \to V$ that satisfy $f(v + w) = f(v) + f(w)$ for all $v, w \in V$, than automorphisms of V as linear space which, in addition, need to satisfy $f(av) = af(v)$ for all $v \in V$ and all $a \in \mathbb{F}$). In fact, the difference is precisely that the latter are all continuous automorphisms of V (considered as topological Abelian group), whereas there are plenty (uncountably many) discontinuous ones, see [28].[31]

Definition 8. Let (S, V, Φ) and (S', V', Φ') be two affine spaces. An **affine morphism** or **affine map** is a pair of maps $F : S \to S'$ and $f : V \to V'$, where f is linear, such that

$$F \circ \Phi = \Phi' \circ f \times F\,. \tag{4.113}$$

[31] Let $\mathbb{F} = \mathbb{R}$, then it is easy to see that $f(v + w) = f(v) + f(w)$ for all $v, w \in V$ implies $f(av) = af(v)$ for all $v \in V$ and all $a \in \mathbb{Q}$ (rational numbers). For continuous f this implies the same for all $a \in \mathbb{R}$. All discontinuous f are obtained as follows: let $\{e_\lambda\}_{\lambda \in I}$ be a (necessarily uncountable) basis of $\mathbb{R}$ as vector space over $\mathbb{Q}$ ('Hamel basis'), prescribe any values $f(e_\lambda)$, and extend f linearly to all of $\mathbb{R}$. Any value-prescription for which $I \ni \lambda \mapsto f(e_\lambda)/e_\lambda \in \mathbb{R}$ is not constant gives rise to a non $\mathbb{R}$-linear and discontinuous f. Such f are "wildly" discontinuous in the following sense: for *any* interval $U \subset \mathbb{R}$, $f(U) \subset \mathbb{R}$ is dense [28].

In the convenient way of writing introduced above, this is equivalent to

$$F(q+v) = F(q) + f(v)\,, \tag{4.114}$$

for all $q \in S$ and all $v \in V$. (Note that the $+$ sign on the left refers to the action Φ of V on S, whereas that on the right refers to the action Φ' of V' on S'.) This shows that an affine map F is determined once the linear map f between the underlying vector spaces is given and the image q' of an arbitrary point q is specified. Equation (4.114) can be rephrased as follows:

Corollary 2. *Let (S, V, Φ) and (S', V', Φ') be two affine spaces. A map $F : S \to S'$ is affine iff each of its restrictions to lines in S is affine.*

Setting $p := q + v$ Eq. (4.114) is equivalent to

$$F(p) - F(q) = f(p-q) \tag{4.115}$$

for all $p, q \in S$. In view of the alternative definition of affine spaces suggested by Proposition 7, this shows that we could have defined affine maps alternatively to (4.113) by ($\Delta' : S' \times S' \to V'$ is the difference map in S')

$$\Delta' \circ F \times F = f \circ \Delta\,. \tag{4.116}$$

Affine bijections of an affine space (S, V, Φ) onto itself form a group, the *affine group*, denoted by $\mathsf{GA}(S, V, \Phi)$. Group multiplication is just given by composition of maps, that is $(F_1, f_1)(F_2, f_2) := (F_1 \circ F_2\,,\ f_1 \circ f_2)$. It is immediate that the composed maps again satisfy (4.113).

For any $v \in V$, the map $F = \Phi_v : p \mapsto p + v$ is an affine bijection for which $f = \mathrm{id}_V$. Note that in this case (4.113) simply turns into the requirement $\Phi_v \circ \Phi_w = \Phi_w \circ \Phi_v$ for all $w \in V$, which is clearly satisfied due to V being a commutative group. Hence there is a natural embedding $T : V \to \mathsf{GL}(S, V, \Phi)$, the image $T(V)$of which is called the subgroup of *translations*. The map $F \mapsto F_* := f$ defines a group homomorphism $\mathsf{GA}(S, V, \Phi) \to \mathsf{GL}(V)$, since $(F_1 \circ F_2)_* = f_1 \circ f_2$. We have just seen that the translations are in the kernel of this map. In fact, the kernel is equal to the subgroup $T(V)$ of translations, as one easily infers from (4.115) with $f = \mathrm{id}_V$, which is equivalent to $F(p) - p = F(q) - q$ for all $p, q \in S$. Hence there exists a $v \in V$ such that for all $p \in S$ we have $F(p) = p + v$.

The quotient group $\mathsf{GA}(S, V, \Phi)/T(V)$ is then clearly isomorphic to $\mathsf{GL}(V)$. There are also embeddings $\mathsf{GL}(V) \to \mathsf{GA}(S, V, \Phi)$, but no canonical one: each one depends on the choice of a reference point $o \in S$, and is given by $\mathsf{GL}(V) \ni f \mapsto F \in \mathsf{GA}(S, V, \Phi)$, where $F(p) := o + f(p-o)$ for all $p \in S$. This shows that $\mathsf{GA}(S, V, \Phi)$ is isomorphic to the semi-direct product $V \rtimes \mathsf{GL}(V)$, though the isomorphism depends on the choice of $o \in S$. The action of $(a, A) \in V \rtimes \mathsf{GL}(V)$ on $p \in S$ is then defined by

$$\big((a, A)\,,\ p\big) \mapsto o + a + A(p-o)\,, \tag{4.117}$$

which is easily checked to define indeed an (o dependent) action of $V \rtimes \mathsf{GL}(V)$ on S.

Appendix 4: Affine Frames, Active and Passive Transformations

Before giving the definition of an affine frame, we recall that of a linear frame:

Definition 9. A **linear frame** of the n-dimensional vector space V over $\mathbb{F}$ is a basis $f = \{e_a\}_{a=1\cdots n}$ of V, regarded as a linear isomorphism $f : \mathbb{F}^n \to V$, given by $f(v^1, \cdots, v^n) := v^a e_a$. The set of linear frames of V is denoted by $\mathcal{F}_V$.

Since $\mathbb{F}$ and hence $\mathbb{F}^n$ carries a natural topology, there is also a natural topology of V, namely that which makes each frame-map $f : \mathbb{F}^n \to V$ a homeomorphism.

There is a natural right action of $\mathsf{GL}(\mathbb{F}^n)$ on $\mathcal{F}_V$, given by $(A, f) \to f \circ A$. It is immediate that this action is simply transitive. It is sometimes called the *passive interpretation* of the transformation group $\mathsf{GL}(\mathbb{F}^n)$, presumably because it moves the frames – associated to the observer – and not the points of V.

On the other hand, any frame f induces an isomorphism of algebras $\text{End}(\mathbb{F}^n) \to \text{End}(V)$, given by $A \mapsto A^f := f \circ A \circ f^{-1}$. If $A = \{A^b_a\}$, then $A^f(e_a) = A^b_a e_b$, where $f = \{e_a\}_{a=1\cdots n}$. Restricted to $\mathsf{GL}(\mathbb{F}^n) \subset \text{End}(\mathbb{F}^n)$, this induces a group isomorphism $\mathsf{GL}(\mathbb{F}^n) \to \mathsf{GL}(V)$ and hence an f-dependent action of $\mathsf{GL}(\mathbb{F}^n)$ on V by linear transformations, defined by $(A, v) \mapsto A^f v = f(Ax)$, where $f(x) = v$. This is sometimes called the *active interpretation* of the transformation group $\mathsf{GL}(\mathbb{F}^n)$, presumably because it really moves the points of V.

We now turn to affine spaces:

Definition 10. An **affine frame** of the n-dimensional affine space (S, V, Φ) over $\mathbb{F}$ is a tuple $F := (o, f)$, where o is a base point in S and $f : \mathbb{F}^n \to V$ is a linear frame of V. F is regarded as a map $\mathbb{F}^n \to S$, given by $F(x) := o + f(x)$. We denote the set of affine frames by $\mathcal{F}_{(S,V,\Phi)}$.

Now there is a natural topology of S, namely that which makes each frame-map $F : \mathbb{F}^n \to S$ a homeomorphism.

If we regard $\mathbb{F}^n$ as an affine space $\text{Aff}(\mathbb{F})$, it comes with a distinguished base point o, the zero vector. The group $\mathsf{GA}\big(\text{Aff}(\mathbb{F}^n)\big)$ is therefore naturally isomorphic to $\mathbb{F}^n \rtimes \mathsf{GL}(\mathbb{F}^n)$. The latter naturally acts on $\mathbb{F}^n$ in the standard way, $\Phi : ((a, A), x) \mapsto \Phi((a, A), x) := A(x) + a$, where group multiplication is given by

$$(a_1, A_1)(a_2, A_2) = (a_1 + A_1 a_2\,,\ A_1 A_2)\,. \tag{4.118}$$

The group $\mathbb{F}^n \rtimes \mathsf{GL}(\mathbb{F}^n)$ has a natural right action on $\mathcal{F}_{(S,V,\Phi)}$, where $(g, F) \mapsto F \cdot g := F \circ g$. Explicitly, for $g = (a, A)$ and $F = (o, f)$, this action reads:

$$F \cdot g = (o, f) \cdot (a, A) = (o + f(a), f \circ A)\,. \tag{4.119}$$

It is easy to verify directly that this is an action which, moreover, is again simply transitive. It is referred to as the *passive interpretation* of the affine group $\mathbb{F}^n \rtimes \mathsf{GL}(\mathbb{F}^n)$.

Conversely, depending on the choice of an affine frame $F \in \mathcal{F}_{(S,V,\Phi)}$, there is a group isomorphism $\mathbb{F}^n \rtimes \mathsf{GL}(\mathbb{F}^n) \to \mathsf{GA}(S, V, \Phi)$, given by $(a, A) \mapsto F \circ (a, A) \circ$

F^{-1}, and hence an F dependent action of $\mathbb{F}^n \rtimes \mathsf{GL}(\mathbb{F}^n)$ by affine maps on (S, V, Φ). If $F = (o, f)$ and $F(x) = p$, the action reads

$$\big((a, A), p\big) \mapsto F(Ax + a) = A^f(p - o) + o + f(a)\,. \tag{4.120}$$

This is called the *active interpretation* of the affine group $\mathbb{F}^n \rtimes \mathsf{GL}(\mathbb{F}^n)$.

An affine frame (o, f) with $f = \{e_a\}_{a=1\cdots n}$ defines $n+1$ points $\{p_0, p_1, \cdots p_n\}$, where $p_0 := o$ and $p_a := o + e_a$ for $1 \leq a \leq n$. Conversely, any $n + 1$ points $\{p_0, p_1, \cdots p_n\}$ in affine space, for which $e_i := p_i - p_0$ are linearly independent, define an affine frame. Note that this linear independence does not depend on the choice of p_0 as our base point, as one easily sees from the identity

$$\sum_{a=1}^{m} v^a(p_a - p_0) = \sum_{k\neq a=0}^{m} v^a(p_a - p_k)\,, \quad \text{where} \quad v^0 := -\sum_{a=1}^{m} v^a\,, \tag{4.121}$$

which holds for any set $\{p_0, p_1, \cdots, p_m\}$ of $m + 1$ points in affine space. To prove it one just needs (4.111). Hence we say that these points are *affinely independent* iff, e.g., the set of m vectors $\{e_a := p_a - p_0 \mid 1 \leq a \leq m\}$ is linearly independent. Therefore, an affine frame of n-dimensional affine space is equivalent to $n + 1$ affinely independent points. Such a set of points is also called an *affine basis*.

Given an affine basis $\{p_0, p_1, \cdots, p_n\} \subset S$ and a point $q \in S$, there is a unique n-tuple $(v_1, \cdots, v_n) \in \mathbb{F}^n$ such that

$$q = p_0 + \sum_{a=1}^{n} v^a(p_a - p_0)\,. \tag{4.122a}$$

Writing $v^k(p_k - p_0) = (p_k - p_0) + (1 - v^k)(p_0 - p_k)$ for some chosen $k \in \{1, \cdots, n\}$ and $v^a(p_a - p_0) = v^a(p_a - p_k) - v^a(p_0 - p_k)$ for all $a \neq k$, this can be rewritten, using (4.111), as

$$q = p_k + \sum_{k\neq a=0}^{n} v^a(p_a - p_k)\,, \quad \text{where} \quad v^0 := 1 - \sum_{a=1}^{n} v^a\,. \tag{4.122b}$$

This motivates writing the sums on the right hand sides of (4.122) in a perfectly symmetric way without preference of any point p_k:

$$q = \sum_{a=0}^{n} v^a p_a,\,, \quad \text{where} \quad \sum_{a=0}^{n} v^a = 1\,, \tag{4.123}$$

where the right hand side is defined by any of the expressions (4.122). This defines certain *linear combinations* of affine points, namely those whose coefficients add up to one. Accordingly, the *affine span* of points $\{p_1, \cdots, p_m\}$ in affine space is defined by

$$\text{span}\{p_1, \cdots, p_n\} := \left\{ \sum_{a=1}^{m} v^a p_a \mid v^a \in \mathbb{F}\,, \sum_{a=1}^{m} v^a = 1 \right\} . \tag{4.124}$$

References

1. Alexandrov, A.D.: Mappings of spaces with families of cones and space-time transformations. Annali di Matematica (Bologna) **103**(8), 229–257 (1975)
2. Bacry, H., Lévy-Leblond, J.-M.: Possible kinematics. J. Math. Phys. **9**(10), 1605–1614 (1968)
3. Beckman, F., Quarles, D.: On isometries of euclidean spaces. Proc. Am. Math. Soc. **4**, 810–815 (1953)
4. Beltrametti, E., Cassinelli, G.: The Logic of Quantum Mechanics. Encyclopedia of Mathematics and its Application, vol. 15. Addison-Wesley, Reading, MA (1981)
5. Berger, M.: Geometry, vol. II, Ist edn. Springer-Verlag, Berlin (1987) Corrected second printing 1996
6. Berger, M.: Geometry, vol. I, Ist edn. Springer-Verlag, Berlin (1987) Corrected second printing 1994
7. Berzi, V., Gorini, V.: Reciprocity principle and the Lorentz transformations. J. Math. Phys. **10**(8), 1518–1524 (1969)
8. Borchers, H.-J., Hegerfeld, G.: The structure of space-time transformations. Commun. Math. Phys. **28**, 259–266 (1972)
9. Born, M.: Die Theorie des starren Elektrons in der Kinematik des Relativitätsprinzips. Annalen der Physik (Leipzig) **30**, 1–56 (1909)
10. Brown, H.: Physical Relativity: Space-Time Structure from a Dynamical Perspective. Oxford University Press, Oxford (2005)
11. Brown, H., Pooley, O.: Minkowski space-time: A glorious non-entity. In: Dieks, D. (ed.). The ontology of spacetime, vol. 1 of Philosophy and foundations of physics, pp. 67–92 (2006)
12. Cantor, G.: Beiträge zur Begründung der transfiniten Mengenlehre. (Erster Artikel). Mathematische Annalen **46**, 481–512 (1895)
13. Casini, H.: The logic of causally closed spacetime subsets. Classic. Quan. Grav. **19**, 6389–6404 (2002)
14. Cegła, W., Jadczyk, A.: Logics generated by causality structures. covariant representations of the Galilei group. Rep. Math. Phys. **9**(3), 377–385 (1976)
15. Cegła, W., Jadczyk, A.: Causal logic of Minkowski space. Commun. Math. Phys. **57**, 213–217 (1977)
16. Coleman, R.A., Korte, H.: Jet bundles and path structures. J. Math. Phys. **21**(6), 1340–1351 (1980)
17. Dixon, W.G.: Special Relativity. The Foundation of Marcroscopic Physics. Cambridge University Press, Cambridge (1978)
18. Ehlers, J., Köhler, E.: Path structures on manifolds. J. Math. Phys. **18**(10), 2014–2018 (1977)
19. Ehrenfest, P.: Gleichförmige Rotation starrer Körper und Relativitätstheorie. Physikalische Zeitschrift **10**(23), 918 (1909)
20. Fock, V.: The Theory of Space Time and Gravitation. Pergamon Press, London (1959)
21. Frank, P., Rothe, H.: Über die Transformation der Raumzeitkoordinaten von ruhenden auf bewegte Systeme. Annalen der Physik (Leipzig) **34**(5), 825–855 (1911)
22. Frank, P., Rothe, H.: Zur Herleitung der Lorentztransformation. Physikalische Zeitschrift **13**, 750–753 (1912) Erratum: ibid, p. 839
23. Frege, G.: Über das Trägheitsgesetz. Zeitschrift für Philosophie und philosophische Kritik **98**, 145–161 (1891)

24. Giulini, D.: Uniqueness of simultaneity. Brit. J. Philos. Sci. **52**, 651–670 (2001)
25. Giulini, D.: Das Problem der Trägheit. Philosophia Naturalis **39**(2), 843–374 (2002)
26. Goldstein, N.J.: Inertiality implies the Lorentz group. Math. Phys. Electro. J. **13**, paper 2 (2007) Available at ⟨www.ma.utexas.edu/mpej/⟩
27. Haag, R.: Local Quantum Physics: Fields, Particles, Algebras. Texts and Monographs in Physics. Springer-Verlag, Berlin (second revised and enlarged edition) (1996)
28. Hamel, G.: Eine Basis aller Zahlen und die unstetigen Lösungen der Funktionalgleichung: $f(x + y) = f(x) + f(y)$. Mathematische Annalen **60**, 459–462 (1905)
29. Herglotz, G.: Über den vom Standpunkt des Relativitätsprinzips aus als 'starr' zu bezeichnenden Körper. Annalen der Physik (Leipzig) **31**, 393–415 (1910)
30. Ignatowsky, W.v.: Einige allgemeine Bemerkungen zum Relativitätsprinzip. Verhandlungen der Deutschen Physikalischen Gesellschaft **12**, 788–796 (1910)
31. Jacobson, N.: Basic Algebra I, 2nd edn. W.H. Freeman, New York (1985)
32. Jafari, N., Ahmad: Operational indistinguishability of varying speed of light theories. Int. J. Mod. Phys. D **13**(4), 709–716 (2004)
33. Jauch, J.M.: Foundations of Quantum Mechanics. Addison-Wesley, Reading, MA (1968)
34. Klein, F.: Vergleichende Betrachtungen über neuere geometrische Forschungen, 1st edn. Verlag von Andreas Deichert, Erlangen (1872) Reprinted in *Mathematische Annalen (Leipzig)* **43**, 43–100 (1892)
35. Lange, L.: Über das Beharrungsgesetz. Berichte über die Verhandlungen der königlich sächsischen Gesellschaft der Wissenschaften zu Leipzig; mathematisch-physikalische Classe **37**, 333–351 (1885)
36. Laue, M.v.: Zur Diskussion über den starren Körper in der Relativitätstheorie. Physikalische Zeitschrift **12**, 85–87 (1911) Gesammelte Schriften und Vorträge (Friedrich Vieweg & Sohn, Braunschweig, 1961), Vol. I, p 132-134
37. Magueijo, J., Smolin, L.: Lorentz invariance with an invariant energy scale. Phys. Rev. Lett. **88**(19), 190403 (2002)
38. Manida, S.N.: Fock-Lorentz transformations and time-varying speed of light. Online available at ⟨http://arxiv.org/pdf/gr-qc/9905046⟩
39. Minkowski, H.: Raum und Zeit. Verlag B.G. Teubner, Leipzig and Berlin, 1909. Address delivered on 21st of September 1908 to the 80th assembly of german scientists and physicians at Cologne
40. Noether, F.: Zur Kinematik des starren Körpers in der Relativitätstheorie. Annalen der Physik (Leipzig) **31**, 919–944 (1910)
41. Pfister, H.: Newton's first law revisited. Found. Phys. Lett. **17**(1), 49–64 (2004)
42. Pirani, F., Williams, G.: Rigid motion in a gravitational field. *Séminaire JANET* (Mécanique analytique et Mécanique céleste) 5e année(8), 1–16 (1962)
43. Robb, A.A.: Optical Geometry of Motion: A New View of the Theory of Relativity. W. Heffer & Sons Ltd., Cambridge (1911)
44. Tait, P.G.: Note on reference frames. Proc. Roy. Soc. Edin. **XII**, 743–745 (Session 1883–84)
45. Thomson, J.: On the law of inertia; the principle of chronometry; and the principle of absolute clinural rest, and of absolute rotation. Proc. Roy. Soc. Edin. **XII**, 568–578 (Session 1883-84)
46. Trautman, A.: Foundations and current problems of general relativity. In: Trautman, A., Pirani, F., Bondi, H. (eds.). Lectures on general relativity, vol. 1 of Brandeis Summer Institute in Theoretical Physics, pp. 1–248. Prentice-Hall, Englewood Cliffs, NJ (1964)
47. Varičak, V.: Über die nichteuklidische Interpretation der Relativtheorie. Jahresberichte der Deutschen Mathematikervereinigung (Leipzig), **21**, 103–127 (1912)
48. Wightman, A.S., Glance, N.: Superselection rules in molecules. Nuc. Phys. B (Proc. Suppl.) **6**, 202–206 (1989)
49. Zeeman, E.C.: Causality implies the Lorentz group. J. Math. Phys. **5**(4), 490–493 (1964)

Chapter 5
Minkowski Space-Time and Quantum Mechanics

W.G. Unruh

Abstract Minkowski's conception of spacetime has had a large impact on our interpretation of quantum theory. Time in quantum mechanics plays a a role in the interpretation distinct from space, in contrast with the apparent unity of space and time encapsulate in the notion of spacetime. It clearly ruled out any quantum mechanics and showed that quantum field theory was the only possible consistent theory. And within quantum field theory, it also rules out the realistic interpretation of the wave function. Fortunately, it is not incompatible with all approaches to quantum theory. Whether or not it is compatible with quantum theory once the metric, the structure which defines the causal structure of the primitive notion of the manifold, is still an open question.

Keywords Minkowski spacetime · Quantum mechanics · Schroedinger representation · Heisenberg representation · Locality · Bell's theorem

5.1 Time, Space, and Quantum Mechanics

In 1908, Minkowski pointed out that Einstein's theory of relativity could be interpreted as if the stage on which physics took place was not space, with changes taking place in space as time proceeded, but rather spacetime, with a metric, a distance function, generalising the notion of distances in space. That is, spacetime behaves like an extended notion of space. But this creates a tension. In common use, even in physics, the notions of space and time differ. One can create plots, in which one of the axes is time and three are space, and plot the evolution of apostrophe one's physical system. However, time is typically believed to have additional structure over space. Time is divided into past, present and future. The past is fixed – it cannot be different than what it is without entering the world of counterfactuals. The

W.G. Unruh (✉)
CIAR Cosmology and Gravity Program, Department of Physics, and Astronomy,
University of B. C., Vancouver, Canada V6T 1Z1
e-mail: unruh@physics.ubc.ca

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_5,

future is ill defined, open with potentialities and possibilities, any of which could be realised in the world. The present is where the possibilities become actual, and where the will of people can exert control over the development of the world. Space, on the other hand, contains no such additional structures. If time becomes absorbed into space via spacetime, this special nature of time becomes much more difficult to uphold. What is the past and the future. Which surface in spacetime acts as the present, and are all possible "presents" equivalent? Or does the notion of a special structure for time simply disappear?

Certainly in a classical deterministic theory, one can argue that the special properties of time are illusory. The history is completely contained in the state of the system at any one time. Nothing can develop because nothing can be different that what it is. The notion that the future is pregnant with possibility is false, it is as fixed and immutable as is the past. The illusion of time's-special-properties becomes a potentially interesting problem, but one of biology or psychology, not physics. Things simply are, in time just as in space.

The problem, of course, is that this is incompatible with the fundamental notions of quantum theory. In quantum mechanics, the universe is not deterministic. Things happen without cause or at least without sufficient cause. The electron passes through this slit or that, not because of something contained in its past, but for no reason whatsoever. Thus the world really does develope in time. The quantum world truly is one of possibilities, only some of which are realised. While the probabilities, in the absence of those actual realisations, are deterministic, the world is not probabilities but actual events and happenings. How can this be reconciled with the Minkowskian world view?

The dominant description of this process is via the Schroedinger picture, in which the dynamic variables of the theory are designated by fixed operators, while it is the state of the system which is dynamic, and changes. However, there are two sorts of change – the one due to the usual dynamic evolution of the state via unitary transformations, and the other, the necessary change in the state when new information about the system is obtained. Since at a time t, under unitary evolution, the state makes only probabilistic statements about the value of some dynamical variable, the determination of that value (for example by a measurement) must be reflected by a change of state, since at that time one and only one value of that variable is actually realised. This second type of change, the collapse of the wave-function, has been regarded as problematic within the world view advocated by Minkowski.

In particular, the problem is usually stated in the form of a question – If the wave function collapses – i.e., changes its values at different points in spacetime, along which of the infinite number of possible definitions of "the same time" does it collapse along. Since any space-like surface (surfaces in which all tangent vectors to curves in that surface have spacelike lengths) could be a special surface defining an instant of time, which are the priviledged ones? To put the problem in the boldest fashion – Is quantum mechanics inconsistent with the Minkowskian view of the world.?

Fortunately this issue vanishes if one examines the problem in more detail. I will go through an number of arguments to show this.

5.2 Heisenberg Representation

The Schroedinger representation is, of course, not the only way in which quantum mechanics can be formulated. In the Heisenberg representation, it is not the state which changes in time, it is the dynamical variables. The unitary transformations change the operators in a time dependent manner. Under the dynamical evolution, the state remains the same. On obtaining extra information about the world, the state changes, as it acts as the aspect of the theory which represents the knowledge (regarded as true knowledge, rather than illusion or personal whim) which we have about the world. When that knowledge changes, the state changes as well, as it is the reflection of that knowledge. Thus the state can no longer be regarded as living in spacetime, as the Schroedinger state for a single particle can. It is simply that differing aspects of knowledge about the world imply different representations of that knowledge within the theory.

Of course this rapidly leads to a further question – namely if I have an attribute of a physical system represented by $A(t)$, a Hermitian operator, and I have bits of knowledge about the world obtained by an interaction with the system at two different times (two different timelike separated points in spacetime), which, if any of those states am I to use in making statements about the attribute $\mathcal{A}$? Simple examples immediately show that a single state is insufficient to encapsulate the knowledge I have about the world, and the impact that knowledge has on my predictions about $\mathcal{A}$. Consider a spin 1/2 particle, about which I have the knowledge that at 9 A.M. it had a value of $+1/2$ for the x component of the spin, and the knowledge that at 11 A.M. it had a value of $+1/2$ for the y component? Furthermore I know that the free Hamiltonian for the particle is zero. How do I incorporate the knowledge about the two values into the theory? While for times earlier than 9 A.M. or later than 11 A.M., I can represent that knowledge by a single state vector (the $+1/2$ eigenstate of S_x before 9 and the $+1/2$ eigenstate of S_y after 11), there exists no state vector which could represent the knowledge that I have about the system at 10 A.M. for example. Had I measured the x component at 10, the prior measurement at 9 would demand that the outcome at 10 be 1/2. Had I measured the y component at 10, the later outcome would demand that I get the value 1/2 for the y component of the spin. But of course there is no state which would predict both certainty for the y and the x components at 10 A.M.

This led Aharonov [3] and collaborators to introduce the idea of the two component wave function, one traveling back into the past from the future, and the other into the future from the past. Both wave functions are then crucial for determining the outcome of intermediate experiments.

$$\mathcal{P}_a = N_A |\langle \Psi_{past}| P_a |\Psi_{future}\rangle|^2$$

where P_a is the usual projection operator onto the eigenstate of A with eigenvalue a. The normalization is not universal but depends on which particular attribute is being considered

$$N_A = \sum_a |\langle \Psi_{past}| P_a |\Psi_{future}\rangle|^2$$

This formula still leaves one with a similar question to the original one. How does one define "future" and how "past"? Minkowski's insight was that these are ambiguous for any two spacelike separated events.

This formula was generalised by a number of people, most recently by GellMan and Hartle [4]. If we consider a sequence of measurements of attributes represented by operators A_i, and we wish to know what the impact on a sequence of future measurements $B_j(t)$ is of the knowledge we have about the $A_i(t)$, they suggest that the appropriate mathematical item is the so called decoherence functional

$$D(\{a_i\},\{b_j\}) = trace(\rho|[P_{a_1}..P_{a_r}P_{b_1}P_{a_{r+1}}...P_{b_2}...]|^2) \tag{5.1}$$

where the order of the projection operators P is in temporal order from past to future (assuming that the initial conditions represented by the density matrix ρ is in the past), and the notation $|C|^2$ for an operator C means $CC^\dagger$ the Hermitian square of the operator. Note that here C is **not** necessarily Hermitian.

It is easy to see that, because in quantum field theory, we believe that any two operators representing attributes concentrated at spacelike separated points commute, the temporal ordering for spacelike separated points in the decoherence functional does not matter. Thus the temporal ordering problem (which of the infinite variety of possibilities for temporal ordering should we take in defining the density functional) vanishes. Any ordering which preserves the causal relations between the operators (i.e., operators defined at causally related times – timelike separated – are ordered) gives the same decoherence functional.

Thus the probability, given the knowledge that we have about the operators A_i at the requisite times, for the values of B_j is just

$$\mathcal{P}_{\{b_j\}} = \frac{D(\{a_i\},\{b_j\})}{\sum_{\{b_j\}} D(\{a_i\},\{b_j\})}$$

Again we have that the normalization factor is not universal – it depends on the attributes B_j that one is measuring.

This emphasizes once again the epistemic nature of the "state". There is not some object (e.g., the Schroedinger wave function which lives out there in the universe) which determines the probabilities of the outcomes of the determinations of the values of various attributes. Instead there is a procedure which takes the knowledge we have about the world, and calculates the probabilities of various outcomes for other determinations that we make.

5.3 Relation Between Quantum Field Theories and Quantum Mechanics

The discussion of the impact of Minkowski's world view on quantum mechanics often takes place in the context of single particle quantum theory. In particular, the Schroedinger wave function for a single particle in the position representation $\Psi(t,\vec{x})$ looks very similar to the expression for some physical quantity $\phi(t,\vec{x})$ defined on the spacetime points. If one is not careful, this leads to succumbing to the temptation of regarding both to be essentially the same.

However, it became obvious during the 1930s that consistency with relativity was impossible for a theory of the quantum mechanics of particles. Not least in leading to this conclusion was the very different character of the the time and the spatial coordinates in this wave function. The spatial coordinates are eigenvalues of an operator, the position operator, while the time is a universal parameter, not associated with any attribute of the particle. The two symbols represent entirely different things. This becomes especially clear if one looks at the wave funtion for two particles, $\Psi(t,\vec{x},\vec{y})$. $\vec{x}$ are the eigenvalues for the position of one particle and $\vec{y}$ are the eigenvalues for the other. There are not two times t associated with the two particles.

It also became rapidly clear to physicists that even a particle based classical mechanics was problematic. The only Lorentz invariant interaction between the two particles (at least those for which a notion of conserved total energy and momentum could be defined) is a contact interaction. Only if the interaction occurs only when the two particles are at the same spacetime point could the system be Lorentz invariant and conserve energy and momentum.

Following the example of the Electromagnetic field, Dirac and Klein-Gordon showed how particles could be replaced by fields. Wigner classified the possible field theories in terms of the representations of the Poincare group. Point interactions now became easy. Since all fields are always defined at all points in spacetime, the demand of locality imposed by Lorentz invariance could be easily satisfied by demanding that the interactions between the fields occur only for fields at the same spacetime point. That is, the interaction between two fields, $\psi(t,x)$ and $\psi(t,x)$ was represented by a product of the fields only at the same spacetime point $\psi(t,x)\psi(t,x)$.

In these fields, the temporal argument and the spatial argument once again represented the same thing– namely the spacetime point at which the field was defined. We note that the Heisenberg picture is also the most natural. It is the quantum operator ϕ which depends on t, not the the "wave function." The wave function, the state, is again time independent (or can be, since physicists often work in the interaction representation, in which the free fields follow their Heisenberg evolution, while the interactions are represented by the unitary evolution of the state.)

5.3.1 Particles

This of course raises the question of the relation between these quantum field theories and single particles. So strong is the pull of the particle picture, that, while physicists always work with quantum fields, they almost invariably describe the physics using particle language. This is taken to the absurd when, for example, the force between two charges is described as the exchange of virtual photons, when all calculations show that it is due to the Coulomb field of the one particle on the other. (I have argued that Feyman diagrams, which are usually pointed to as support for the particle picture, set back physics by 20 years, because it blinded physicists to what are called non-perturbative effects– effects which are often trivial when looked at from the field point of view.) In particular almost all cogitations about the foundations of quantum mechanics are in terms of particles, localized in certain regions of spacetime, not in terms of fields. The reason this is possible of course is that one can often treat the quantum mechanics of a quantum field in terms of single particles, and as if that particle had a definite wave function defining the probabilities for those particles.

Let us take a quantum field theory, with the "free" quantum field Φ, where I will assume that the value of the field at some point in spacetime is let us say two valued. That is, we could write this field Φ in terms of two fields $\tilde{\Phi}_1$ and $\tilde{\Phi}_2$. Let the vacuum state for this field be $|0\rangle$. We can decompose this quantum field into normal modes ϕ_i where the index i is infinite. (again ϕ_i is actually two valued at any position in spacetime.) In general I do not want these to be plane wave modes, but relatively localized modes. These modes will obey the equation

$$\Box\phi_i + m^2\phi_i = 0 \tag{5.2}$$

One of the requirements of the mode $\phi_i(t, x)$ is that it have positive norm.

$$< \phi_i, \phi_j >= \frac{i}{2}\int \left[\phi_i^\dagger \partial_\mu \phi_j - \partial_\mu \phi_i^\dagger \phi_j\right] dS_\mu = \delta_{ij} \tag{5.3}$$

where dS_μ is the three volume form on the spacelike hyper-surface over which the integration is being done. Because the field obeys the wave equation, this integral is independent of the surface on which it is defined, and remains normalized. If $\{\phi_i\}$ and $\{\phi_i^*\}$ are a complete orthonormal set of positive and negative norm modes, (i.e., such that if $< \phi_i, \tilde{\phi} >= 0$ for all i, then $< \tilde{\phi}, \tilde{\phi} >< 0$) and the quantum field is designated by Φ, then $a_i =< \phi_i, \Phi >$ is an annihilation operator for the field, obeying $a_j, a_j^\dagger >= \delta_{ij}$. One can define the "vacuum" state by $|i\rangle = a_i|0\rangle = 0$, and a one particle state by $a_i^\dagger|0\rangle$, or a normalized sum of these states ($\sum_i \alpha_i a_i^\dagger|0\rangle$ with $\sum_i |\alpha_i|^2 = 1$).

Then the function ϕ_i acts in many ways like the wave function for a single particle. We could define a particle detector by placing a large a two level system in a region of space, switched on for some period of time, such that it interacts with the field in a region around x and absorbs the particle (e.g., something that binds the electron, or absorbs the photon). The interaction between the detector and the field looks like

$$\left[\int \Phi(x,t)h(x,t)\sigma_1 d^3x dt + E\sigma_3\right] \tag{5.4}$$

where σ_i are the usual Pauli spin matrices controlling the two level system representing the detector. The function h(x, t) is the coupling parameter between the field and the detector, and is assumed to be smooth in both x and t. and to incorporate the coupling constant. The energy E of the detector determines which energy this particular detector responds to . Note that I will assume that h is small and keep only terms to lowest order in h.

Note that this model is one in which the even if the particle is in the region, the probability is small that it is actually detected. Thus one has to look at the relative probabilities – how does the probability depend on the amount of the field within the region of interest.

The answer is that the probability that the detector is excited is

$$\mathcal{P} = \left| \int \phi_i(t,x) h(t,x) e^{-iEt} d^3x dt \right|^2 \tag{5.5}$$

Assuming the ϕ_i has a temporal oscillation of the order of E but is smooth in space over the extent of $h(x,t)$ this corresponds to having the detection probability being proportional to $i \int |\phi_i(t,x)|^2 h(t,x) d^3x dt$ which is just what we would expect if were were trying to calculate the probability of finding a particle with wave function ϕ_i in the region delimited by $h(x)$. That is, under certain conditions, the field theory representation of the system can be approximated by an effective Schroedinger representation, where, instead of the magnitude of the field at any point in spacetime being the observable, the position of the particle becomes the observable. This means that in a certain approximate sense, one can still talk as if the Schroedinger description of particles were the correct description.

Note however that the equation obeyed by the field itself, the true observable, is completely local and relativistic. It obeys the field equation

$$\Box\Phi + m^2\Phi = 0 \tag{5.6}$$

If any perturbation is introduced into the field Φ by, for example, a local coupling to some other field, the effects of that perturbation are felt only within the future null cone of the original perturbation region. There is nothing about Quantum Field theory, which, as I argued, is the only possible viable relativistic theory, which is in conflict with the Minkowskian world view.

The important point is that it is not the position of the particle that is detected. It is some function of the field strength of the particle in a certain spacetime region which is detected. If that function is such that if the particle were completely contained in that region (ϕ_i were non-zero only within that region during the detection time), then the detection probability would be unity, then one can regard that measurement of the field strength as if it were a detection of the particle in that region. We note however, that the measurement is one of field within a region, and does not, except for the correlations implicit in the initial state, imply anything about the field strengths in any other region. The measurement process is completely local.

(The above is of course only a sketch of how quantum field strengths can, under certain conditions, play the role of the wavefunction for a particle. Details will be presented elsewhere.)

5.4 Locality

At this point I can hear in my mind's ear a chorus chanting "But everyone knows that quantum mechanics is non-local, and non-locality is surely a violation of the Minkowski world view". I will finish this paper by arguing that this is wrong. Quantum theory is not non-local. This misconceptions arises out of a misunderstanding of Bell's famous theory. In fact I will argue that the use of locality in that theorem is to make the classical system behave as much like the quantum system as possible. "Locality" plays the role, not of differentiating classical from quantum, but rather in making the classical system as close to the quantum as possible.

Let us go through Bell's theorem (as modified by Clauser, Horne and Shimony [2]), first crudely and then in a more detailed manner. Bell postulates two particles, two localized physical entities, well separated in space. Each of these particles has some set of attributes, A, B for the first particle and C, D for the second. Each of these attributes, in conformity with the discrete nature of quantum mechanics, is assumed to have possible values of ± 1. The actual values assumed in any particular realisation of the experiment are assumed to be random. Furthermore in conformance with quantum mechanics, it is assumed that one cannot, in the same experiment actually measure (determine the value of) both A and B. One could postulate that there is something in the measurement of say A which will interfere with the measurement of B, so that the very process of determining A could alter the value of B and vice verso. The same is true for the attributes C and D of the second particle.

We now imagine a series of experiments, in each of which the two particles are placed into the same physical state, time after time. What this state really encompasses, we will leave vague, but the state is such that the probabilities of the outcomes of any determination of any attribute or attributes are the same in each trial. The physical processes which create these two states are as identical in each of the runs as they can be made.

One now imagines a series of four different types of experiments. These could be selected by some arbitrary random process, the randomness being determined by whatever source one desired. In particular the random choice can be made separately at each of the particles. In each random choice at the first particle either A or B is determined. In each random choice at the second particle, either C or D is determined. Thus there are four possible series of determinations selected by this process.

In one series of experiments, one determines the values of A for the first particle and of C for the second. One takes the average over all trials of the product of the values of A and C. Let us call this average $< AC >$ In another series, one determines A and D, keeping again the initial state the same, and can then calculate the average $< AD >$. In the third series one determines $< BC >$ and in the fourth $< BD >$.

One now examines the following sum of all of these averages

$$\mathcal{G} \equiv < AC > - < AD > + < BC > + < BD > \tag{5.7}$$

The next step is the critical step, which we will come back to later, the step in which one relies on the various locality and other postulates. Each of the averages is determined in a separate series of trials chosen at random. One now argues that the above expression can be written as

$$\mathcal{G} = < AC - AD + BC + BD > \tag{5.8}$$

where the average on the right hand side is assumed be evaluated in each individual instance of the run. In the situation envisaged by Bell, the right hand side is taken to mean that if, with a god's eye view of the experiment, one could know what the values of A, B, C, and D were, and one could evaluate this expression in each individual instance and then take the average over a large number of runs. That is, in no individual instance is this sum of attributes ever determined. It is in that sense only a hypothetical average. It is establishing this equality which is the most difficult part of the Bell argument. Much has been written as to the assumptions one needs to make this statement, but locality, the fact that the first particle and the second can be taken to inhabit space-like separated regions in space-time, always plays a crucial role.

I will not review these arguments, as they will turn out to be largely irrelevant. I will simply assume that the appropriate arguments have been made, that appropriate appeals to locality, to parameter independence, to completeness, or whatever other assumptions, have been made and we have established this equality. We can now write the right hand side of this expression as

$$< AC - AD + BC + BD > = < (A + B)C + (B - A)D > \tag{5.9}$$

Since A and B have possible values of ± 1 the value of $A + B$ is ± 2 or 0. Similarly for $A - B$. Furthermore the value of $A + B$ is correlated with that of $B - A$ in that anytime $A + B$ has value 0, $B - A$ has value ± 2 and vice verso. Since C and D have values of ± 1 this implies that on each run of the experiment, the value of $(A + B)C + (B - A)D$ is ± 2 and thus the average lies between -2 and $+2$. Thus we have

$$-2 < \mathcal{G} < 2 \tag{5.10}$$

which is the Clauser-Horne-Shimony form of Bell's theorem.

The key feature of this argument that people have most discussed is Eq. 5.8. Under what circumstances can one genuinely argue that that the sum of the averages, carried out in separate experiments, is the same as the average of the sum of the values of the attributes in any single experiment averaged over all trials? As stated, it is here that locality plays its crucial role. The alternative is that each of the correlations could depend on the measurements made. If the values determined for A, for example, or the probabilistic distribution for the values of A, depended on whether C or D were measured at the other particle, then clearly the equivalence could be violated. That the two particles can be regarded as spatially separated and that the whole experiment could take place on a time scale much smaller than that required for the any information traveling at the speed of light could travel between the two. Thus the distribution of the probabilities of the values of A must surely be independent of which of C or D is determined at the other particle. Similarly for B, and similarly for C and D with respect to whether A or B were determined.

Since these argument seem so clear-cut and since quantum mechanics violates the bound given by Bell for this sum of correlations, it has become common to say that Quantum Mechanics must violate locality. This is however a perverse conclusion because quantum mechanics satisfies exactly this requirement given in Eq. 5.1.

Let me make specific the quantum situation and state. We choose the two attributes A and B of the first particle to be represented by the two σ matrices σ_1 and σ_3. The two eigenstates of σ_1 will be designated by $|+\rangle$ and $|-\rangle$. For the second particle, we choose C and D to be two different Σ matrices associated with this particle,

$$C = \frac{1}{\sqrt{2}}\left(\Sigma_1 + \Sigma_3\right) \tag{5.11}$$

$$D = \frac{1}{\sqrt{2}}\left(-\Sigma_1 + \Sigma_3\right) \tag{5.12}$$

Again the eigenstates of Σ_1 are designated by $|+\rangle$ and $|-\rangle$. We will choose the total state to be given by

$$|\psi\rangle = \frac{1}{\sqrt{2}}\left(|+\rangle|+\rangle + |-\rangle|-\rangle\right) \tag{5.13}$$

This is the maximally correlated state for any determination of any of the operators $X = cos(\theta)\sigma_1 + sin(\theta)\sigma_3$ with $Y = cos(\theta)\Sigma_1 + sin(\theta)\Sigma_3$. That is, anytime X has value $+1$ so does Y and anytime X has value -1 so does Y. (Note that I have taken this rather than the usual total spin $= 0$ maximally anti-correlated state.)

It is now easy in this state to determine the correlations of measurements of A and C, A and D, B and C, and B and D. The sum $\mathcal{G}$ is $2\sqrt{2}$, which is larger than 2, and in violation of Bell's theorem.

However, we recall that the difficult step in the proof of Bell's theorem was to argue that $\mathcal{G} =< AC - AD + BC + BD >$. But quantum mechanics gives us this for free. That is, such linearity of expectation values is contained in the fundamental postulates of quantum mechanics. It is trivial to show that in any state $|\Psi\rangle$ the operators A,B,C and D obey the operator equation

$$\langle\Psi|XY|\Psi\rangle + \langle\Psi|WZ|\Psi\rangle = \langle\Psi|(XY + WZ)|\Psi\rangle \tag{5.14}$$

It is precisely in this step that one had to use locality, etc., in the classical argument, a trivial conclusion in quantum mechanics. That is, quantum mechanics and classical mechanics – with all of the locality, parameter independence, etc – agree completely on this relationship.

In the quantum case one does not really want the equivalence of Eqs. 5.7 and 5.8. The resultant operator, $AC_AD + BC + BD$ is a non-local operator, which cannot be determined by individual measurements at the individual location of the particle. However, this is not required. Instead we write

$$\mathcal{G} =< (A + B)C > + < (B - A)D > \tag{5.15}$$

where now each of the averages, $< (A + B)C >$ and $< (B - A)D >$ can be determined using a separate series of trials in which only attributes at the two particles need be measured.

Then where does quantum mechanics differ from classical? Where is it that quantum mechanics comes to violate Bell's inequality? The answer is that the operators corresponding to $A + B$ and $A - B$ do not behave as the Bell argument says they behave. They do not have values of $\pm 2,\ 0$. They are attributes in their own right (in both quantum mechanics and classical mechanics) and they have values of $\pm\sqrt{2}$. Furthermore, $A + B$ and $A - B$ are not anti-correlated as the classical argument would have them be. That is, for any value of $A + B$ one can say nothing about the value of $B - A$, except that they are distributed 50-50 over the two possible values of $B - A$. However, $A + B$ is completely correlated with C and $B - A$ with D. Thus each of the two terms in the sum $< (A + B)C + (B - A)D >$ has the same value of $\sqrt{2}$, leading to the quantum result of $2\sqrt{2}$.

It is of course true that in the quantum case, the equivalence of $\mathcal{G}$ and the sum means something different than it does in the classical case. In the classical case, Bell interpreted $< AC - AD + BC + BD >$ as the correlation function in which the values corresponding to each of A, B, C, D are definite in any one trial, and this expectation value is the average over all trials of those products of individual values. In the quantum case, $AC - AD + BC + BD$ is an operator corresponding to the whole system. It is furthermore an operator which cannot be regarded as any kind of statement about any one of the systems at a time. In the classical case of course, one could argue, with the same assumptions which gives the Bell result, that

$$\mathcal{G} =< (A + B)C > + < (B - A)D > \tag{5.16}$$

However, Bell's explicit assumption is that each of these terms is not the same as in the quantum case. That is, over a large series of trials, these averages of of each of the two terms will be given by the quantum result, namely that $A + B$ will be perfectly correlated with C (and $B - A$ with D) and that furthermore, the values obtained for $A + B$ or $A - B$ will not be $\pm\sqrt{2}$ but rather would be $\pm 2,\ 0$. This is not a statement about locality but a statement about classical mechanics not being able to replicate quantum mechanics on the single particle level, or being able to replicate the simple complete correlation of the two observables.

Bell's further step in arguing that $(A + B)^2$ and $(A - B)^2$ are "anti-correlated" in any one trial, which is the final step leading to the Bell limit, is really only the counterfactual icing on the cake. It is already the individual correlations $< (A + B) C >$ and $< (B - A)D >$ which do not agree with quantum mechanics,

We thus see that the focus of most of the concern in the comparison of the quantum system to the classical has focused around arguing the equivalence of the two averages. But it is not in that equivalence that quantum and classical systems differ. It is in the fact that the values which the sum of two attributes can have differs from the sum of the values that the individual attributes. Even if A and B each have value ± 1, the sum $A + B$ has values $\pm\sqrt{2}$. This is of course a well known aspect of quantum mechanics, and has nothing to do with locality.

Bell's theorem certainly shows that quantum and classical systems differ, but the difference really resides in the local and well known property of quantum mechanics, that the "sum of the values" of two attributes is not the same as the "value of the sum" of those attributes, while they are equal in the classical realm.

5.4.1 Leggett's Extention of Bell's Theorem for Non-Linear Classical System

Leggett suggested that a large class of of non-local theories, which however obey conditions on the single particle statistics, also obey inequalities similar to Bell's for local theories. He called these theories crypto-non-local theories. Instead of providing a general characterisation of these theories, he built a specific example.

The example he developed was to describe a hidden variable theory for some experiments on the coherent creation of polarized light from a source. This source produces two "beams" of light (two photons) which one measures the polarization of. As in the Aspect type experiments, he is concerned with the correlations between the polarization of the two photons.

The experiment has a source which produces the two photons, which head toward two detectors. These detectors measure the polarization of the light the light of the two photons, and from these measurements one calculates the correlation coefficients for the various choices of orientation of the two detectors.

I will work in the Stokes parameter formalism. The normalized Stokes parameters are characterized by a three dimensional vector, the components representing vertical-horizontal polarization, left-right diagonal polarization, and left and right circular polarization. The primary assumption Leggett makes is that in any particular experiment, the photon always has a particular direction of polarization, which I will designate as $\vec{u}$ and $\vec{v}$ for the photons heading toward the two detectors.

The detectors themselves are characterized by a similar three dimensional vector to designate which polarization they are designed to detect. It is assumed, in conformity with experimental evidence, that on entering the detector, the outcome of the experiment is either a $+1$, to designate the fact that that detector has measured the photon to have the polarization which the detector is designed to measure, and a -1 if it does not. Furthermore he assumes that for any one photon, the decision whether it has that polarization or not is a binary process. That is, the outcome of the experiment on any single photon is always either ± 1.

The detectors are characterized by a unit three dimensional vector, $\vec{a}$ and $\vec{b}$ for the two detectors to represent the direction of polarization that that detector is designed to measure. The components of the vectors represent the three Stokes parameters of the measuring apparatus.

In any single experiment, there are assumed to be a set of "hidden variables". These are the polarization parameters, $\vec{u}\ \vec{v}$ of the two photons, and an extra hidden variable λ_{uv} which will determine, given the polarization of the photons, what the interaction with the measuring apparatus of those photons will be. That is, given $\vec{u}$, $\vec{v}$ and λ_{uv} the interaction with any measuring apparatus or apparatuses is determined. There is however no assumption of locality in these.

The source of the two photons is then assumed to be characterized by two probability distributions, $\rho(\lambda_{uv})$, the probability distribution over λ_{uv} for given $\vec{u}\ \vec{v}$, and $F(\vec{u}, \vec{v})$, the probability distribution over the polarization of the two photons. Both of these probability distributions are assumed to be fixed by the source, and to be independent of the setting of, or existence of, the measuring apparatuses. That is, there is an assumption of locality in time (the source occurring earlier in time than the measurements, the measurements cannot have an effect on the source).

Now, given the settings on the measuring apparatuses, $\vec{a}\ \vec{b}$, the outcomes of the measurements are designated by A and B. These represent the outcome of the measurement, and both have values of ± 1. The value of the outcome is assumed to be determined by the values of the settings of the measuring apparatuses, $\vec{a}\ \vec{b}$, by the values of the polarization of the two photons $\vec{u}\ \vec{b}$, and by the value of the parameter λ_{uv}. However, no assumption is made about the function $A(\vec{a},\ \vec{b},\ \vec{u},\ \vec{v},\ \lambda_{uv})$ and $B(\vec{a},\ \vec{b},\ \vec{u},\ \vec{v},\ \lambda_{uv})$, except for one restriction. In particular, no assumption is made that A depends only on $\vec{a}$ and $\vec{u}$ for example. The value can depend on $\vec{v}$ and $\vec{b}$ as well. That is, the outcome of the experiment can be non-local.

However, Leggett does place one experimentally motivated restriction on the dependence of A on its arguments. Given $\vec{u}$ and $\vec{v}$, the average value of A on its own should follow the known dependence of the average of the Stokes parameters on the polarization. Namely he demands that

$$\bar{A} = \int A((\vec{a}, \vec{b}, \vec{u}, \vec{v}, \lambda_{uv})\rho(\lambda_{uv})d\lambda_{uv} = \vec{a}\cdot\vec{u} \tag{5.17}$$

That is, the average value of A given the polarization of the photons obeys the experimentally known distribution for the polarization measurements. Similarly we have

$$\bar{B} = \vec{b}\cdot\vec{v} \tag{5.18}$$

where $\vec{B}$ is defined in the same way.

Let us define the following expressions for given values of $\vec{u}\ \vec{v}\ \vec{a}\ \vec{b}$

$$\rho_{++} = \int \delta(A-1)\delta(B-1)\rho(\lambda_{uv})d\lambda_{uv}d\lambda_{uv} \tag{5.19}$$

$$\rho_{+-} = \int \delta(A-1)\delta(B+1)\rho(\lambda_{uv})d\lambda_{uv}d\lambda_{uv} \tag{5.20}$$

$$\rho_{-+} = \int \delta(A+1)\delta(B-1)\rho(\lambda_{uv})d\lambda_{uv}d\lambda_{uv} \tag{5.21}$$

$$\rho_{-+} = \int \delta(A+1)\delta(B+1)\rho(\lambda_{uv})d\lambda_{uv}d\lambda_{uv} \tag{5.22}$$

Then we have

$$\begin{aligned}
-1+|\bar{A}+\bar{B}| &= -(\rho_{++}+\rho_{+-}+\rho_{-+}+\rho_{--})+|(\rho_{++}+\rho_{+-}-\rho_{-+}-\rho_{--}) \\
&\quad +(\rho_{++}+\rho_{-+}-\rho_{+-}-\rho_{--})| && (5.23)\\
&= -(\rho_{++}+\rho_{+-}+\rho_{-+}+\rho_{--})+2|\rho_{++}-\rho_{--}| && (5.24)\\
&\leq -(\rho_{++}+\rho_{+-}+\rho_{-+}+\rho_{--})+2(\rho_{++}+\rho_{--}) && (5.25)\\
&= \rho_{++}+\rho_{--}-\rho_{+-}-\rho_{-+} = \int AB\rho(\lambda_{uv})d\lambda_{uv} = \overline{AB} && (5.26)
\end{aligned}$$

We now average over the possible values of the internal orientation, namely over $\vec{u}$, $\vec{v}$, and use the crypto-non-local condition, to get

$$\int F(\vec{u},\vec{v})(-1+|\vec{a}\cdot\vec{u}+\vec{b}\cdot\vec{v}|d^2\vec{u}d^2\vec{v} \leq \int F(\vec{u},\vec{v})\bar{A}\bar{B}d^2\vec{u}d^2\vec{v} =< AB > \tag{5.27}$$

where $< AB >$ is the expectation value of the correlation in the experiment of the product of the values of A and B, and is measurable in a series of experiments.

Now, let us assume that the source is the quantum source such that the emission of the photons place them into the singlet state. This means that quantum mechanically, $< AB > = -\vec{a}\cdot\vec{b}$. If our crypto-non-local theory is to mimic quantum mechanics, we must choose our distribution function $F(\vec{u},\vec{v})$ to obey

$$\int F(\vec{u},\vec{v})(-1+|\vec{a}\cdot\vec{u}+\vec{b}\cdot\vec{v}|) \leq -\vec{a}\cdot\vec{b} \tag{5.28}$$

If we choose $\vec{a}=\vec{b}$, then we must have

$$\int F(\vec{u},\vec{v})|\vec{a}\cdot(\vec{u}+\vec{v})|d^2ud^2v = 0 \tag{5.29}$$

for all values of $\vec{a}$. This implies that

$$F(\vec{u},\vec{v}) = \delta(\vec{u}+\vec{v})\mathcal{F}(\vec{u}) \tag{5.30}$$

Now, we define

$$\vec{a}-\vec{b} = 2\sin\left(\frac{\psi}{2}\right)\vec{e} \tag{5.31}$$

where $\vec{a}\cdot\vec{b} = \cos(\psi)$.

Thus in order that the classical hidden variables theory mimic the quantum, we must have that

$$-1+2\int \mathcal{F}(\vec{u})\sin\left(\frac{\psi}{2}\right)|\vec{e}\cdot\vec{u}|d^{u} \leq -cos(\psi) \tag{5.32}$$

or

$$2sin\left(\frac{\psi}{2}\right)\left(sin\left(\left(\frac{\psi}{2}\right)-\int \mathcal{F}(\vec{u})|\vec{e}\cdot\vec{u}|d^{2}u\right) \geq 0 \tag{5.33}$$

for all unit vectors $\vec{e}$ and all angles ψ. Clearly Eq. (5.33) can be satisfied only if

$$\int \mathcal{F}(\vec{u})|\vec{e}\cdot\vec{u}|d^{2}u = 0 \tag{5.34}$$

for all vectors $\vec{e}$. Since $\mathcal{F}(\vec{u}) \geq 0$ and $\int \mathcal{F}(\vec{u})d^{2}u = 1$, Eq. 5.34 is impossible to satisfy.

Averaging the expression over $\vec{e}$, we have

$$\frac{1}{\int d^{2}e}\int\left(2\sin\left(\frac{\psi}{2}\right)\left(sin\left(\frac{\psi}{2}\right)-\int \mathcal{F}(\vec{u})|\vec{e}\cdot\vec{u}|d^{2}u\right)\right)d^{2}e \tag{5.35}$$

$$= \sin\left(\frac{\psi}{2}\right)\left(sin\left(\frac{\psi}{2}\right)-\frac{1}{2}\int \mathcal{F}(\vec{u})d^{2}u\right) \tag{5.36}$$

$$= 2\sin\left(\frac{\psi}{2}\right)\left(\sin(\psi)-\frac{1}{2}\right) \tag{5.37}$$

which is minimized when $sin(\psi) = 1/4$ with a value of $-\frac{1}{8}$.

Also, since the integrand $\int \mathcal{F}(\vec{u})|\vec{e}\cdot\vec{u}|d^{2}u$ is positive for all values of $\vec{e}$, then if we maximize the integrand over $\vec{e}$, that maximum must be larger than or equal to the average. Thus

$$\mathrm{Max}_{\vec{e}}\int \mathcal{F}(\vec{u})|\vec{e}\cdot\vec{u}|d^{2}u \geq \frac{1}{2} \tag{5.38}$$

Thus there must always exist a value of $\vec{e}$ for which the violation of the inequality is greater than $\frac{1}{8}$.

That is, for any distribution $F(\vec{u},\vec{v})$, the maximum violation of the classical inequality over all vectors $\vec{e}$ or equivalently of $\vec{a}$ and $\vec{b}$ is bounded below by 0.125. No classical crypto-non-local theory can mimic the correlations of quantum mechanics. That is, even the presence of non-locality in space cannot make a classical system equivalent to a quantum system.

5.5 Conclusion

We have examined a variety of ways which people have naively used to argue that quantum mechanics is somehow inconsistent with the Minkowski vision of the structure of space and time. In each case, I would argue that the arguments are invalid, and at the very least that one can interpret quantum theory in a way that is consistent.

These arguments say nothing about the impact of quantum gravity on the vision. It is of course true that the much more complex structure of General Relativity already alters the causal structure of Minkowski's spacetime. The simplest case, pointed out by Penrose[1] almost 40 years ago, is that in flat spacetime two spatially separated observers can always escape from each other. One simply travels at the velocity of light directly away from the other. The other can never catch up.

On the other hand, if one is closer to a mass than the other, then that other can always catch up. The light cone of the further observer always completely encloses the light cone of first sufficiently far into the future. Essentially because of the Shapiro time delay, light rays which curve around the mass always get ahead of a light ray directed directly away from the first observer.

This is in addition to the well known features of General Relativity in which one can have closed time-like curves in certain spacetimes.

In quantum gravity, because of general covariance, when interpreted as quantum constraints, seem to imply that locality looses all meaning. While clearly wrong, the solution to this puzzle is still a problem for the future.

References

1. Bell, J.S.: Physics **1**, 195 (1964)
2. Clauser, J.F., Horne, M.A., Shimony, A., Holt, R.A.: Phys. Rev. Lett. **23**, 880 (1969)
3. Reznik, B., Aharonov, Y.: Phys. Rev. A **52**, 2538 (1995) [arXiv:quant-ph/9501011]
4. Gell-Mann, M., Hartle, J.B.: SFI Studies in the Sciences of Complexity. In: Zurek, W. (ed.) Complexity, entropy and the physics of information, vol. VIII. Addison Wesley, Reading (1990)
5. Unruh, W.: Int. J. Quan. Inform. (*IJQI*), (This section is largely a rewrite of a part of the paper) **4**, 209–218 (2006)
6. Leggett, A.: Found. Phys. **33**, 1469 (2003)

[1] He pointed this out to me in private conversations, but I have not found a suitable reference. The observation was that while in flat spacetime, the future light cone of any point never completely encompasses the intersection of the future light cone of another spacelike separated point with a timelike surface, if one point is closer to a black hole, or any other gravitating mass, then the future light cone of the further point always completely encompasses the intersection of the future light cone of the nearer point with a timelike surface sufficiently far into the future. I.e., any observer located at the point nearer the mass can never escape the blast wave originating at the further point, even if light like velocities are possible.

Chapter 6
Modern Space-Time and Undecidability

Rodolfo Gambini and Jorge Pullin

Abstract The picture of space-time that Minkowski created in 1907 has been followed by two important developments in physics not contained in the original picture: general relativity and quantum mechanics. We will argue that the use of concepts of those theories to construct space-time implies conceptual modifications in quantum mechanics. In particular one can construct a viable picture of quantum mechanics without a reduction process that has outcomes equivalent to a picture with a reduction process. One therefore has two theories that are entirely equivalent experimentally but profoundly different in the description of reality they give. This introduces a fundamental level of undecidability in physics of a kind that has not been present before. We discuss some of the implications.

Keywords Decoherence · Measurement problem · Undecidability

6.1 Introduction

In 1907 Minkowski noted that the natural arena to formulate the special theory of relativity of Einstein was space-time. In his address delivered at the 80th Assembly of German Natural Scientists and Physicians in 1908, Minkowski remarks: "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." "The soil of experimental physics" in 1907 differs from the one today. At that time it was thought that space and time were continuous and that they could be measured with arbitrary

R. Gambini
Institute of Physics, Montevideo University, Uruguay

J. Pullin
Horace Hearne Jr. Institute for Theoretical Physics Center for Computation and Technology
Department of Physics & Astronomy Louisiana State University, Baton Rouge, LA 70803-4001
e-mail: pullin@lsu.edu

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_6,

precision. This was altered when quantum mechanics was developed in the 1930s. There it was noted that fundamental uncertainties were germane to modern physics. In 1957 Salecker and Wigner [1] decided to revise the picture that Minkowski had painted of space-time introducing concepts of quantum mechanics. In particular they considered limitations in the accuracy of clocks and rulers that one needs in order to construct the picture of space-time that we are familiar with. But fundamental uncertainties in quantum mechanics can be minimized. The situation changes dramatically when general relativity is brought into the picture in two different ways. On the one hand general relativity leads to fundamental uncertainties that cannot be minimized. In particular one cannot measure distances and times beyond a minimum level of uncertainty, as was emphasized by Ng and Van Dam [2]. The second way in which general relativity alters the picture is that space-time is now a dynamical arena that is not directly observable. The only observable properties of space-time are relational in nature. Therefore unlike the space-time of Minkowski, which was an immutable arena for special relativity, in general relativity space-time becomes an object that cannot be directly probed and the properties of it that can be probed are relational in nature and subject to fundamental minimum uncertainties.

To further elaborate the point, consider the measurement of time. In its usual formulation, quantum mechanics involves an idealization. The idealization is the use of a perfect classical clock to measure times. Such a device clearly does not exist in nature, since all measuring devices are subject to some level of quantum fluctuations. Therefore the equations of quantum mechanics, when cast in terms of the variable that is really measured by a clock in the laboratory, will differ from the traditional Schrödinger description. Although this is an idea that arises naturally in ordinary quantum mechanics, it is of paramount importance when one is discussing quantum gravity. This is due to the fact that general relativity is a generally covariant theory where one needs to describe the evolution in a relational way. One ends up describing how certain objects change when other objects, taken as clocks, change. At the quantum level this relational description will compare the outcomes of measurements of quantum objects. Quantum gravity is expected to be of importance in regimes (e.g., near the big bang or a black hole singularity) in which the assumption of the presence of a classical clock is clearly unrealistic. The question therefore arises: is the difference between the idealized version of quantum mechanics and the real one just of interest in situations when quantum gravity is predominant, or does it have implications in other settings?

We will argue that indeed it does have wider implications. Some of them are relevant to conceptual questions (e.g., the black hole information paradox or ultimate limitations on quantum computing) and there might even be experimental implications. A detailed discussion of these ideas can be found in previous papers [3–5], and in particular in the pedagogical review [6]. Here we present an abbreviated discussion as an introduction to a remarkable consequence of these ideas: that the nature of physical processes in modern physics becomes *undecidable*. In a nutshell we observe that when one considers fundamental limitations in the measurements of space-time, unitary quantum mechanics behaves in the same way as quantum mechanics with a reduction process. The two theories therefore become equivalent in their physical predictions, yet they are radically different in nature: in one of them quantum states are determined once and for all at the beginning of the universe and are not changed by physical events. In the other the picture is dramatically different, with quantum states constantly changing as events produces measurable outcomes. In spite of these differences, there is no experimental way to decide which of these pictures corresponds to reality. The implications of this observation philosophically are profound. They advocate what is technically known in philosophical circles as a "regularist" point of view towards the nature of physical laws, drastically different from the "necessitarian" point of view that was considered at the time of Minkowski's conception of space-time as the most natural interpretation of the role of physical laws.

The plan of this paper is as follows: in the next section we will discuss the form of the evolution equation of quantum mechanics when the time variable, used to describe it, is measured by a real clock. In Section 6.3 we will consider a fundamental bound on how accurate can a real clock be and the implications it has for quantum mechanics in terms of real clocks and its consequences. Section 6.4 discusses briefly some possible experimental implications of the proposal. Section 6.5 discusses conceptual implications in the foundations of quantum mechanics. Section 6.6 argues that the conceptual implications leads to a new notion of undecidability in the nature of physical processes in quantum theory. We end with a discussion.

6.2 Quantum Mechanics with Real Clocks

Given a physical situation we start by choosing a "clock". By this we mean a physical quantity (more precisely a set of quantities, like when one chooses a clock and a calendar to monitor periods of more than a day) that we will use to keep track of the passage of *time*. An example of such a variable could be the angular position of the hand of an analog watch. Let us denote it by T. We then identify some physical variables that we wish to study as a function of time. We shall call them generically O ("observables"). We then proceed to quantize the system by promoting all the observables and the clock variable to self-adjoint quantum operators acting on a Hilbert space. The latter is defined once a well defined inner product is chosen in the set of all physically allowed states. Usually it consists of squared integrable functions $\psi(q)$ with q the configuration variables.

Notice that we are not in any way modifying quantum mechanics. We assume that the system has an evolution in terms of an external parameter t, which is a classical variable, given by a Hamiltonian and with operators evolving with Heisenberg's equations (it is easier to present things in the Heisenberg picture, though it is not mandatory to use it for our construction). Then the standard rules of quantum mechanics and its probabilistic nature apply.

We will call the eigenvalues of the "clock" operator T and the eigenvalues of the "observables" O. We will assume that the clock and the measured system do not interact (if one considered an interaction it would produce additional effects to the one discussed). So the density matrix of the total system is a direct product of that of the system under study and the clock $\rho = \rho_{\rm sys} \otimes \rho_{\rm cl}$, and the system evolves through a unitary evolution operator that is of the tensor product form $U = U_{\rm sys} \otimes U_{\rm cl}$. The quantum states are described by a density matrices at a time t. Since the latter is unobservable, we would like to shift to a description where we have density matrices as functions of the observable time T. We define the probability that the resulting measurement of the clock variable T correspond to the value t,

$$\mathcal{P}_t(T) \equiv \frac{\mathrm{Tr}\left(P_T(0)U_{\rm cl}(t)\rho_{\rm cl}U_{\rm cl}(t)^\dagger\right)}{\int_{-\infty}^{\infty} dt\, \mathrm{Tr}\left(P_T(t)\rho_{\rm cl}\right)}, \tag{6.1}$$

where $P_T(0)$ is the projector on the eigenspace with eigenvalue T evaluated at $t = 0$. We note that $\int_{-\infty}^{\infty} dt \mathcal{P}_t(T) = 1$. We now define the evolution of the density matrix,

$$\rho(T) \equiv \int_{-\infty}^{\infty} dt U_{\rm sys}(t)\rho_{\rm sys}U_{\rm sys}(t)^\dagger \mathcal{P}_t(T) \tag{6.2}$$

where we dropped the "sys" subscript in the left hand side since it is obvious we are ultimately interested in the density matrix of the system under study, not that of the clock.

We have therefore ended with an "effective" density matrix in the Schrödinger picture given by $\rho(T)$. It is possible to reconstruct entirely in a relational picture the probabilities using this effective density matrix, for details we refer the reader to the lengthier discussion in [6]. By its very definition, it is immediate to see that in the resulting evolution unitarity is lost, since one ends up with a density matrix that is a superposition of density matrices associated with different t's and that each evolve unitarily according to ordinary quantum mechanics.

Now that we have identified what will play the role of a density matrix in terms of a "real clock" evolution, we would like to see what happens if we assume the "real clock" is behaving semi-classically. To do this we assume that $\mathcal{P}_t(T) = f(T - T_{\rm max}(t))$, where f is a function that decays very rapidly for values of T far from the maximum of the probability distribution $T_{\rm max}$. We refer the reader to [6] for a derivation, but the resulting evolution equation for the probabilities is (in the limit in which corrections are small),

$$\frac{\partial \rho(T)}{\partial T} = i[\rho(T), H] + \sigma(T)[H, [H, \rho(T)]]. \tag{6.3}$$

and the extra term is dominated by the rate of change $\sigma(T)$ of the width of the distribution $f(t - T_{\rm max}(t))$.

An equation of a form more general than this has been considered in the context of decoherence due to environmental effects, it is called the Lindblad equation. Our particular form of the equation is such that conserved quantities are automatically preserved by the modified evolution. Other mechanisms of decoherence coming from a different set of effects of quantum gravity have been criticized in the past because they fail to conserve energy [7]. It should be noted that Milburn arrived at a similar equation as ours from different assumptions [8]. Egusquiza, Garay and Raya derived a similar expression from considering imperfections in the clock due to thermal fluctuations [9]. It is to be noted that such effects will occur in addition to the ones we discuss here. Corrections to the Schrödinger equation from quantum gravity have also been considered in the context of WKB analyses [10]. Considering time as a quantum variable has also been discussed by Bonifacio [11] with a formulation somewhat different than ours but with similar conclusions.

Another point to be emphasized, particularly in the context of a volume celebrating Minkowski's contributions, is that our approach has been quite naive in the sense that we have kept the discussion entirely in terms of non-relativistic quantum mechanics with a unique time across space. It is clear that in addition to the decoherence effect we discuss here, there will also be decoherence spatially due to the fact that one cannot have clocks perfectly synchronized across space and also that there will be fundamental uncertainties in the determination of spatial positions. This, together with the issue of the Lorentz invariance of the predictions, is discussed in some detail in our paper [12].

6.3 Fundamental Limits to Realistic Clocks

We have established that when we study quantum mechanics with a physical clock (a clock that includes quantum fluctuations), unitarity is lost, conserved quantities are still preserved, and pure states evolve into mixed states. All this in spite of the fact that the underlying theory is unitary as usual. The effects are more pronounced the worse the clock is. Which raises the question: is there a fundamental limitation to how good a clock can be? This question was first addressed by Salecker and Wigner [1]. Their reasoning went as follows: suppose we want to build the best clock we can. We start by insulating it from any interactions with the environment. An elementary clock can be built by considering a photon bouncing between two mirrors. The clock "ticks" every time the photon strikes one of the mirrors. Such a clock, even completely isolated from any environmental effects, develops errors. The reason for them is that by the time the photon travels between the mirrors, the wavefunctions of the mirrors spread. Therefore the time of arrival of the photon develops an uncertainty. Salecker and Wigner calculated the uncertainty to be $\delta t \sim \sqrt{t/M}$ where M is the mass of the mirrors and t is the time to be measured (we are using units where $\hbar = c = 1$ and therefore mass is measured in 1/s). The longer the time measured the larger the error. The larger the mass of the clock, the smaller the error.

So this tells us that one can build an arbitrarily accurate clock just by increasing its mass. However, Ng and Van Dam [2] pointed out that there is a limit to this. Basically, if one piles up enough mass in a concentrated region of space one ends up with a black hole. Some readers may ponder why do we need to consider a concentrated region of space. The reason is that if we allow the clock to be more massive by making it bigger, it also deteriorates its performance (see the discussion in [13] in response to [14]).

A black hole can be thought of as a clock since it is an oscillator. In fact it is the "fastest" oscillator one can have, and therefore the best clock for a given size. It has normal modes of vibration that have frequencies that are of the order of the light travel time across the Schwarzschild radius of the black hole. (It is amusing to note that for a solar sized black hole the frequency is in the kilohertz range, roughly similar to that of an ordinary bell). The more mass in the black hole, the lower the frequency, and therefore the worse its performance as a clock. This therefore creates a tension with the argument of Salecker and Wigner, which required more mass to increase the accuracy. This indicates that there actually is a "sweet spot" in terms of the mass that minimizes the error. Given a time to be measured, light traveling at that speed determines a distance, and therefore a maximum mass one could fit into a volume determined by that distance before one forms a black hole. That is the optimal mass. Taking this into account one finds that the best accuracy one can get in a clock is given by $\delta T \sim T_{\text{Planck}}^{2/3} T^{1/3}$ where $t_{\text{Planck}} = 10^{-44}s$ is Planck's time and T is the time interval to be measured. This is an interesting result. On the one hand it is small enough for ordinary times that it will not interfere with most known physics. On the other hand is barely big enough that one might contemplate experimentally testing it, perhaps in future years.

With this absolute limit on the accuracy of a clock we can quickly work out an expression for the $\sigma(T)$ that we discussed in the previous section [5,15]. It turns out to be $\sigma(T) = \left(\frac{T_{\text{Planck}}}{T_{\max}-T}\right)^{1/3} T_{\text{Planck}}$. With this estimate of the absolute best accuracy of a clock, we can work out again the evolution of the density matrix for a physical system in the energy eigenbasis. One gets n

$$\rho(T)_{nm} = \rho_{nm}(0)e^{-i\omega_{nm}T}\, e^{-\omega_{nm}^2 T_{\text{Planck}}^{4/3} T^{2/3}}. \tag{6.4}$$

So we conclude that *any* physical system that we study in the lab will suffer loss of quantum coherence at least at the rate given by the formula above. This is a fundamental inescapable limit. A pure state inevitably will become a mixed state due to the impossibility of having a perfect classical clock in nature.

6.4 Possible Experimental Implications

Given the conclusions of the previous section, one can ask what are the prospects for detecting the fundamental decoherence we propose. At first one would expect them to be dim. It is, like all quantum gravitational effects, an "order Planck" effect. But it should be noted that the factor accompanying the Planck time can be rather large. For instance, if one would like to observe the effect in the lab one would require that the decoherence manifest itself in times of the order of magnitude of hours, perhaps days at best. That requires energy differences of the order of $10^{10}eV$ in the Bohr frequencies of the system. Such energy differences can only be achieved in "Schrödinger cat" type experiments, but are not outrageously beyond our present capabilities. Among the best candidates today are Bose–Einstein condensates, which can have 10^6 atoms in coherent states. However, it is clear that the technology is still not there to actually detect these effects, although it could be possible in forthcoming years.

A point that could be raised is that atomic clocks currently have an accuracy that is less than a decade of orders of magnitude worse than the absolute limit we derived in the previous section. Couldn't improvements in atomic clock technology actually get better than our supposed absolute limit? This seems unlikely. When one studies in detail the most recent proposals to improve atomic clocks, they require the use of entangled states [16] that have to remain coherent. Our effect would actually prevent the improvement of atomic clocks beyond the absolute limit!

Finally, if one has doubts about the effect's existence, one must recall that one can make it arbitrarily large just by picking a lousy clock. This is of course, not terribly interesting and is not what is normally done in physics labs. But it should be noted that experiments of Rabi oscillations in rubidium atoms measure certain correlations which can be interpreted as having the atom work as a lousy clock. The oscillations show experimentally the exponential decay we discuss. See Bonifacio et al. [17] for a discussion.

6.5 Conceptual Implications

The fact that pure states evolve naturally into mixed states has conceptual implications in at least three interesting areas of physics. The first two we have discussed before so we only mention them for completeness and to refer the reader to previous papers: the black hole information paradox [5, 15] and quantum computation [18]. The third area where the effects we discussed could be of interest are in foundational issues in quantum mechanics, in particular, the measurement problem. We will now expand a bit on what we mean by this.

The measurement problem in quantum mechanics is related to the fact that in ordinary quantum mechanics the measurement apparatus is assumed to be always in an eigenstate after a measurement has been performed. The usual explanation [19] for this is that there exists interaction with the environment. This selects a preferred basis, i.e., a particular set of quasi-classical states that commute, at least approximately, with the Hamiltonian governing the system-environment interaction. Since the form of the interaction Hamiltonians usually depends on familiar "classical" quantities, the preferred states will typically also correspond to the small set of "classical" properties. Decoherence then quickly damps superpositions between the localized preferred states when only the system is considered. This is taken as an explanation of the appearance to a local observer of a "classical" world of determinate, "objective" (robust) properties.

There are two main problems with such a point of view. The first one is how is one to interpret the local suppression of interference in spite of the fact that the total state describing the system-environment combination retains full coherence. One may raise the question whether retention of the full coherence could ever lead to empirical conflicts with the ascription of definite values to macroscopic systems. The usual point of view is that it would be very difficult to reconstruct the off diagonal elements of the density matrix in practical circumstances. However, at least as a matter of principle, one could indeed reconstruct such terms (the evolution of the whole system remains unitary [20]) by "waiting long enough". The second problem is that even if the state ends up being "quasi-diagonal" in the preferred basis of the measurement device, this is not necessarily a completely satisfactory solution to the measurement problem. This is known as the "and-or" problem. As Bell [28] put it "if one were not actually on the lookout for probabilities, ... the obvious interpretation of even ρ' [the reduced density matrix] would be that the system is in a state in which various $|\Psi_m>$'s coexist:

$$|\Psi_1><\Psi_1| \quad \text{and} \quad |\Psi_2><\Psi_2| \quad \text{and} \quad \ldots \tag{6.5}$$

This is not at all a *probability* interpretation, in which the different terms are seen not as *coexisting* but as *alternatives*." We will return on this point when we introduce our idea of undecidability later in this paper.

Our mechanism of fundamental decoherence could contribute to the understanding of the two problems mentioned above. In the usual system-environment interaction the off-diagonal terms of the density matrix oscillate as a function of time. Since the environment is usually considered to contain a very large number of degrees of freedom, the common period of oscillation for the off-diagonal terms to recover non-vanishing values is very large, in many cases larger than the life of the universe. This allows to consider the problem solved in practical terms. When one adds in the effect we discussed, since it suppresses exponentially the off-diagonal terms, one never has the possibility that the off-diagonal terms will see their initial values restored, no matter how long one waits.

To analyze the implications of the use of real clocks in the measurement problem, we will discuss in some detail an example. In spite of the universality of the loss of coherence we introduced, it must be studied in specific examples of increasing level of realism. The simplest example we can think of is due to Zurek [22]. This simplified model does not have all the effects of a realistic one, yet it exhibits how the information is transferred from the measuring apparatus to the environment. The model consists of taking a spin one-half system that encodes the information about the microscopic system plus the measuring device. A basis in its two dimensional Hilbert space will be denoted by $\{|+>, |->\}$. The environment is modeled by a bath of many similar two-state systems called atoms. There are N of them, each denoted by an index k and with associated two dimensional Hilbert space $\{|+>_k, |->_k\}$. The dynamics is very simple, when there is no coupling with the environment the two spin states have the same energy, which is taken to be 0. All the atoms have zero energy as well in the absence of coupling. The whole dynamics is contained in the coupling, given by the following interaction Hamiltonian

$$H_{\text{int}} = \hbar \sum_k \left(g_k \sigma_z \otimes \sigma_z^k \otimes \prod_{j \neq k} I_j \right). \tag{6.6}$$

In this notation σ_z is analogous to a Pauli spin matrix. It has eigenvalues $+1$ for the spin eigenvector $|+>$ and -1 for $|->$; it acts as the identity operator on all the atoms of the environment. The operators σ_z^k are similar, each acts like a Pauli matrix on the states of the specific atom k and as the identity upon all the other atoms and the spin. I_j denotes the identity matrix acting on atom j and g_k is the coupling constant that has dimensions of energy and characterizes the coupling energy of one of the spins k with the system. In spite of the abstract character of the model, it can be thought of as providing a sketchy model of a photon propagating in a polarization analyzer.

Starting from a normalized initial state

$$|\Psi(0)>=(a|+>+b|->)\prod_{k=1}^{N}\otimes\left[\alpha_k|+>_k+\beta_k|->_k\right], \tag{6.7}$$

it is easy to solve the Schroedinger equation and one gets for the state at the time t,

$$\begin{aligned}|\Psi(t)> = a|+>\prod_{k=1}^{N}\otimes\left[\alpha_k\exp\left(ig_kt\right)|+>_k+\beta_k\exp\left(-ig_kt\right)|->\right] \\ +b|->\prod_{k=1}^{N}\otimes\left[\alpha_k\exp\left(-ig_kt\right)|+>_k+\beta_k\exp\left(ig_kt\right)|->\right].\end{aligned} \tag{6.8}$$

Writing the complete density operator $\rho(t)=|\Psi(t)><\Psi(t)|$, one can take its trace over the environment degrees of freedom to get the reduced density operator,

$$\rho_c(t)=|a|^2|+><+|+|b|^2|-><-|+z(t)ab^*|+><-|+z^*(t)a^*b|-><+|, \tag{6.9}$$

where

$$z(t)=\prod_{k=1}^{N}\left[\cos\left(2g_kt\right)+i\left(|\alpha_k|^2-|\beta_k|^2\right)\sin\left(2g_kt\right)\right]. \tag{6.10}$$

The complex number $z(t)$ controls the value of the non-diagonal elements. If this quantity vanishes the reduced density matrix ρ_c would correspond to a totally mixed state ("proper mixture"). That would be the desired result, one would have several classical outcomes with their assigned probabilities. However, although the expression we obtained vanishes quickly assuming the α's and β's take random values, it behaves like a multiperiodic function, i.e. it is a superposition of a large number of periodic functions with different frequencies. Therefore this function will retake values arbitrarily close to the initial value for sufficiently large times. This implies that the apparent loss of information about the non-diagonal terms reappears if one waits a long enough time. This problem is usually called "recurrence of coherence". The characteristic time for these phenomena is proportional to the factorial of the number of involved frequencies. Although this time is usually large, perhaps exceeding the age of the universe, at least in principle it implies that the measurement process does not correspond to a change from a pure to a mixed state in a fundamental way.

The above derivation was done using ordinary quantum mechanics in which one assumes an ideal clock is used to measure time. If one re-does the derivation using the effective equation we derived for quantum mechanics with real clocks one gets the same expression for $z(t)$ except that it is multiplied by $\prod_k \exp\left(-(2g_k)^2 T_{\text{Planck}}^{4/3} t^{2/3}\right)$. That means that asymptotically the off diagonal terms indeed vanish, the function $z(t)$ is not periodic anymore. Although the exponential term decreases slowly with time, the fact that there is a product of them makes the effect quite relevant.

Therefore we see that the inclusion of real clocks makes impossible the observation of "revivals" in systems where the measurement process leads to observable outcomes. The observation of "revivals" not only is very difficult to observe in a practical sense due to the length of time that elapses between revivals, but becomes unobservable due to fundamental reasons, irrespective of the advanced level of the technology used for the observations. As a consequence the revivals are not a means to determine if global coherence in the total system has been preserved or if there has been a change in the state of the measurement system.

Revivals are not the only possible manifestation of the total unitary evolution of the system when an observation of a given result is made in the system. Other experiments have been proposed. These are also of considerable practical difficulty since they involve measurements of the total system. For example, d'Espagnat [23] has proposed a technique based on the observation of constant of the motion of the total system. Such constants of the motion take different values if the reduction process has happened or not. These experiments require the construction of macroscopic ensembles including the environment prepared in the same state and subsequent measurements of each of the degrees of freedom involved. The concrete models proposed [23] for these types of experiments are highly idealized, ignoring interactions between the degrees of freedom of the environment and their individual evolution. The free Hamiltonians of the environment and the measuring apparatus is assumed zero. In these models the loss of unitarity due to real clocks appears to be insufficient to eliminate all possibilities of distinguishing between a reduction and a unitary evolution in an observable process due to the fact that one would be dealing with a constant of the motion. Preliminary analyses of concrete implementations of these ideas in more realistic models lead us to believe that the loss of spatial coherence and dispersion during evolution of the wavefunctions of the component degrees of freedom and the back reaction of the measuring device on the system [24] plus the evolution of the measuring device appear to imply that the observation of the relevant observables is impossible. We are currently studying some of these models and will report elsewhere on the details.

6.6 Undecidability and the Laws of Physics

We have shown that quantum mechanics with and without an explicit mechanism for state reduction yield the same physical predictions. There are good reasons to consider that it is experimentally impossible to distinguished between two alternative situations, the one resulting from a unitary evolution of the system plus the environment and the other one where the system plus the measuring device undergo an abrupt change of state. There are therefore two complete logical structures that yield the same predictions and therefore one cannot decide experimentally which one "corresponds to reality". This is what we refer to as "undecidability".

The concept of undecidability in physics was also discussed by Wolfram [25] in a different context, although not totally disconnected from the one we introduce here. He refers to the undecidability of computational nature about the outcome of a physical process. There exist physical processes whose outcome is not predictable. The optimum computational process to determine the outcome is the physical process itself. In this sense we could speak of *undecidability in the outcome* of a process. This type of undecidability is weaker than the one we are discussing, since the issues can be decided just by waiting till the outcome occurs. The undecidability we are referring to does not refer to the outcome of a process but to the nature of the process.

In philosophy there are different attitudes that have been taken towards the physical laws of nature (see for instance [26]). One of them is the "regularity theory"; in it, the laws of physics are statements about uniformities or regularities of the world and therefore are just "convenient descriptions" of the world. The laws of physics are dictated by a preexisting world and are a representation of our ability to describe the world. Another point of view is the "Necessitarian theory". There laws of nature are "principles" which govern the natural phenomena, that is, the world "obeys" the laws of nature. The presence of the undecidability we point out suggests strongly that the "regularity theory" point of view is more satisfactory.

If one takes seriously the regularity point of view one can ponder about the nature of reality. Does the physical world have a reduction process, does it not, or does it depend on the case? In the case in which there is no reduction process, in the Heisenberg picture the state of a system is given and eternal. If there is a reduction process the state changes every time there is an event resulting in a measurement. The third possibility, which is suggested by the undecidability, is that the system may choose between behaving as if there is a reduction process or not. This would add to the well known probabilistic freedom of quantum systems characterized by the free will theorem of Conway and Kochen [27] another one characterized by two alternative behaviors in relation to the rest of the universe. That is, after the observation of the event either the system simply behaves as if it were part of the universe and its state were that of the universe or if as its state would be given by the reduction postulate. In the first case the system would keep its entanglement with the rest of the universe (i.e., the environment), in the second it will lose its entanglement.

If one adopts what is probably the most attractive view that considers that the universe always evolves unitarily and therefore quantum states are determined once and for all no matter what is the chosen behavior of the subsystem under observation one needs to face the problem of when do events happen in such a framework. Our point of view is that an event occurs when the experimental distinction between coexisting or exclusionary alternatives becomes undecidable, since in that instant the predictions of the laws of physics are not altered by the possible reduction of the state of the system associated with the information acquired when the event takes place.

6.7 Discussion

We have argued that the use of realistic clocks in quantum mechanics implies that pure states evolve into mixed states in spite of the fact that the underlying equations of quantum mechanics are unitary. The effect is further compounded if one allows realistic rulers to measure distances and the formulation can be made Lorentz covariant. This alters fundamentally the picture of space-time as a static fixed entity that was first introduced by Minkowski. Conceptually, it also has profound implications since it allows to construct a quantum mechanics without the introduction of a reduction process that nevertheless has the same predictions as ordinary quantum mechanics with a reduction process. The nature of physical processes is therefore undeterminate in a novel and fundamental way that adds itself to other previous proposals of undecidability in physics.

Acknowledgements This work was supported in part by grants NSF-PHY0650715, and by funds of the Horace C. Hearne Jr. Institute for Theoretical Physics, FQXi, PEDECIBA (Uruguay) and CCT-LSU.

References

1. Wigner, E.: Rev. Mod. Phys. **29**, 255 (1957)
2. Ng, Y.J., Dam, H.v.: Ann. N. Y. Acad. Sci. **755**, 579 (1995) [arXiv:hep-th/9406110]; Mod. Phys. Lett. A **9**, 335 (1994); see also Károlhyázy, F., Frenkel, A., Lukács, B. "Quantum concepts in space and time" Penrose, R., Isham, C., (eds.). Oxford University Press, Oxford (1986)
3. Gambini, R., Porto, R., Pullin, J.: Class. Quant. Grav. **21**, L51 (2004) [arXiv:gr-qc/0305098]
4. Gambini, R., Porto, R., Pullin, J.: New J. Phys. **6**, 45 (2004) [arXiv:gr-qc/0402118]
5. Gambini, R., Porto, R., Pullin, J.: Braz. J. Phys. **35**, 266 (2005) [arXiv:gr-qc/0501027]
6. Gambini, R., Porto, R., Pullin, J.: Gen. Rel. Grav. **39**, 1143 (2007) [arXiv:gr-qc/0603090]
7. Ellis, J., Hagelin, J., Nanopoulos, D.V., Srednicki, M.: Nucl. Phys. **B241** (1984) 381; Banks, T., Peskin, M.E., Susskind, L. Nucl. Phys. **B244** (1984) 125
8. Milburn, G.J.: Phys. Rev **A44**, 5401 (1991)
9. Egusquiza, I., Garay, L., Raya, J.: Phys. Rev. **A59**, 3236 (1999) [arXiv:quant-ph/9811009]
10. Kiefer, C., Singh, T.: Phys. Rev. **D44**, 1061 (1991)
11. Bonifacio, R.: Nuo. Cim. **D114**, 473 (1999)
12. Gambini, R., Porto, R.A., Pullin, J.: Int. J. Mod. Phys. D **15**, 2181 (2006) [arXiv:gr-qc/0611148]
13. Ng, Y.J., Dam, H.v.: Class. Quant. Grav. **20**, 393 (2003) [arXiv:gr-qc/0209021]

14. Baez, J.C., Olson, S.J.: Class. Quant. Grav. **19**, L121 (2002) [arXiv:gr-qc/0201030]
15. Gambini, R., Porto, R.A., Pullin, J.: Phys. Rev. Lett. **93**, 240401 (2004) [arXiv:hep-th/0406260]
16. Andre, A., Sorensen, A., Lukin, M.: Phys. Rev. Lett **92**, 230801 (2004) [arXiv:quant-ph/0401130]
17. Bonifacio R. et al.: Phys. Rev. **A61**, 053802 (2000)
18. Gambini, R., Porto, R., Pullin, J.: in Gravity, astrophysics and strings at the black sea, Fiziev, P., Todorov, M. (eds.). St. Kliment Ohridski Press (2006) [arXiv:quant-ph/0507262]
19. Schlosshauer, M.: Rev. Mod. Phys. **76**, 1267 (2004) [arXiv:quant-ph/0312059]
20. Omnès, R.: The Interpretation of Quantum Mechanics. Princeton Series in Physics, Princeton, NJ (1994)
28. Bell, J.: Against 'Measurement'. In: Miller, A. (ed.). Sixty two years of uncertainty, Plenum, New York (1990)
22. Zurek, W.: Phys. Rev. **D26**, 1862 (1982)
23. d'Espagnat, B.: Veiled Reality. Westview, Boulder, CO (2003)
24. Bartlett, S., Rudolph, T., Spekkens, R., Turner, P.: New J. Phys. **8**, 58 (2006); Poulin, D. Yard, J. New J. Phys. **9** 156 (2007)
25. Wolfram, S.: Phys. Rev. Lett. **54**, 735 (1985)
26. "Laws of Nature" in The Stanford encyclopedia of Philosophy (http://plato.stanford.edu) and references therein
27. Conway, J., Kochen, S.: Found. Phys. **36**, 1441 (2006)

Chapter 7
Quantum Space-Times
Beyond the Continuum of Minkowski and Einstein

Abhay Ashtekar

Abstract In general relativity space-time ends at singularities. The big bang is considered as the Beginning and the big crunch, the End. However these conclusions are arrived at by using general relativity in regimes which lie well beyond its physical domain of validity. Examples where detailed analysis is possible show that these singularities are naturally resolved by quantum geometry effects. Quantum space-times can be vastly larger than what Einstein had us believe. These non-trivial space-time extensions enable us to answer of some long standing questions and resolve of some puzzles in fundamental physics. Thus, a century after Minkowski's revolutionary ideas on the nature of space and time, yet another paradigm shift appears to await us in the wings.

Keywords Space-time Singularities · Big Bang · Black Holes · Information Loss · Singularity Resolution · Quantum Geometry · Quantum Bounce · Space-time Extension

7.1 Introduction

A 100 years ago Hermann Minkowski fused space and time into a smooth four-dimensional continuum. Remarkably, this continuum – the Minkowski space-time – still serves as the arena for all non-gravitational interactions both in classical *and* quantum physics. Time is no more absolute. Whereas in Newtonian physics there is a unique three-plane through each space-time point representing space, now there is a unique cone, spanned by light rays passing through that point. The constant time plane curls up into a two sheeted cone that separates the region which is causally connected with the point from the region which is not. This causality dictates the propagation of physical fields in classical physics, and the commutation relations

A. Ashtekar (✉)
Institute for Gravitation and the Cosmos and Physics Department Penn State, University Park, PA 16802, USA
e-mail: Ashtekar@gravity.psu.edu

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_7,

and uncertainty relations between field operators in quantum physics. With the demise of absolute simultaneity, Newtonian ideas are shattered. The world view of physics is dramatically altered.

However, as in Newtonian physics, there is still a fixed space-time which serves as the arena for all of physics. It is the stage on which the drama of evolution unfolds. Actors are particles and fields. The stage constrains what the actors can do. The Minkowski metric dictates the field equations and restricts the forms of interaction terms in the action. But the actors cannot influence the stage; Minkowskian geometry is immune from change. To incorporate the gravitational force, however, we had to abandon this cherished paradigm. We follow Einstein and encode gravity in the very geometry of space-time. Matter curves space-time. The space-time metric is no longer fixed. There is again a dramatic paradigm shift. However, we continue to retain one basic feature of Newtonian and Minkowskian frameworks: space-time is still represented by a smooth continuum.

This is not uncommon: New paradigms are often created by abandoning one key feature of the older paradigm but retaining another. But global coherence of the description of Nature is a huge burden and such a strategy often leads to new tensions. For example, to achieve compatibility between mechanics and Maxwellian electrodynamics, Einstein abandoned absolute simultaneity but retained the idea that space and time are fixed, unaffected by matter. The strategy worked brilliantly. Not only was the new mechanics compatible with Maxwell's theory but it led to deep, unforeseen insights. Energy and mass are simply two facets of the same physical attribute, related by $E = mc^2$; electric and magnetic fields $\mathbf{E}$, $\mathbf{B}$ are but two projections of an electromagnetic field tensor F_{ab}; in a quantum theory of charged particles, each particle must be accompanied by an anti-particle with opposite charge. However, the new mechanics flatly contradicted basic tenets of Newton's theory of gravitation. To restore coherence of physics, one has to abandon the idea that space-time is fixed, immune to change. One had to encode gravity into the very geometry of space-time, thereby making this geometry dynamical.

Now the situation is similar with general relativity itself. Einstein abandoned the tenet that geometry is inert and made it a physical entity that interacts with matter. This deep paradigm shift again leads to unforeseen consequences that are even more profound. Thanks to this encoding, general relativity predicts that the universe began with a big bang; that heavy stars end their lives through a gravitational collapse to a black hole; that ripples in the space-time curvature propagate as gravitational waves carrying energy-momentum. However, general relativity continues to retain the Newtonian and Minkowskian premise that space-time is a smooth continuum. As a consequence, new tensions arise.

In Newtonian or Minkowskian physics, a given physical field could become singular at a space-time point. This generally implied that the field could not be unambiguously evolved to the future of that point. However, this singularity had no effect on the global arena. Since the space-time geometry is unaffected by matter, it remains intact. Other fields could be evolved indefinitely. Trouble was limited to the one field which became ill behaved. However, because gravity is geometry in general relativity, when the gravitational field becomes singular, the continuum tares and the space-time itself ends. There is no more an arena for other fields to live in. All of physics, as we know it, comes to an abrupt halt. Physical observables associated with both matter and geometry simply diverge signalling a fundamental flaw in our description of Nature. This is the new quandary.

When faced with deep quandaries, one has to carefully analyze the reasoning that led to the impasse. Typically the reasoning is flawed, possibly for subtle reasons. In the present case the culprit is the premise that general relativity – with its representation of space-time as a smooth continuum – provides an accurate description of Nature arbitrarily close to the singularity. For, general relativity completely ignores quantum effects and, over the last century, we have learned that these effects become important in the physics of the small. They should in fact be *dominant* in parts of the universe where matter densities become enormous. Thus there is no reason to trust the predictions of general relativity near space-time singularities. Classical physics of general relativity does come to a halt at the big-bang and the big crunch. But this is not an indication of what *really* happens because use of general relativity near singularities is an extrapolation which has no physical justification whatsoever. We need a theory that incorporates not only the dynamical nature of geometry but also the ramifications of quantum physics. We need a quantum theory of gravity, a new paradigm.

These considerations suggest that singularities of general relativity are perhaps the most promising gates to physics beyond Einstein. They provide a fertile conceptual and technical ground in our search of the new paradigm. Consider some of the deepest conceptual questions we face today: the issue of the Beginning and the end End; the arrow of time; and the puzzle of black hole information loss. Their resolutions hinge on the true nature of singularities. In my view, considerable amount of contemporary confusion about such questions arises from our explicit or implicit insistence that singularities of general relativity are true boundaries of space-time; that we can trust causal structure all the way to these singularities; that notions such as event horizons are absolute even though changes in the metric in a Planck scale neighborhood of the singularity can move event horizons dramatically or even make them disappear altogether [1].

Over the last 2–3 years several classically singular space-times have been investigated in detail through the lens of loop quantum gravity (LQG) [2–4]. This is a non-perturbative approach to the unification of general relativity and quantum physics in which one takes Einstein's encoding of gravity into geometry seriously and elevates it to the quantum level. One is thus led to build quantum gravity using *quantum* Riemannian geometry [5–8]. Both geometry and matter are *dynamical* and described *quantum mechanically* from the start. In particular, then, there is no background space-time. The kinematical structure of the theory has been firmly established for some years now. There are also several interesting and concrete proposals for dynamics (see, in particular [2–4, 9]). However, in my view there is still considerable ambiguity and none of the proposals is fully satisfactory. Nonetheless, over the last 2–3 years, considerable progress could be made by restricting oneself to subcases where detailed and explicit analysis is possible [10–15]. These "mini" and "midi" superspaces are well adapted to analyze the deep conceptual tensions discussed above. For, they consider the most interesting of classically singular space-times – Friedman-Robertson-Walker (FRW) universes with the big bang singularity and black holes with the Schwarzschild-type singularity – and analyze them in detail using symmetry reduced versions of loop quantum gravity. In all cases studied so far, classical singularities are naturally resolved and *the quantum space-time is vastly larger than what general relativity had us believe.* As a result, there is a new paradigm to analyze the old questions.

The purpose of this article is to summarize these developments, emphasizing the conceptual aspects[1] from an angle that, I hope, will interest not only physicists but especially philosophers and historians of science. We will see that some of the long standing questions can be directly answered, some lose their force in the new paradigm while others have to be rephrased.

This chapter is organized as follows. In Section 7.2 I will discuss cosmological singularities and in 7.3 the black hole singularities. In each case I will discuss examples of fundamental open issues and explain their status in the corresponding models. We will see that quantum geometry has unexpected ramifications that either resolve or significantly alter the status of these issues. Finally in Section 7.4 I will summarize the outlook and discuss some of the fresh challenges that the new paradigm creates.

7.2 Quantum Nature of the Big Bang

7.2.1 Issue of the Beginning and the End

Over the history of mankind, cosmological paradigms have evolved in interesting ways. It is illuminating to begin with a long range historical perspective by recalling paradigms that seemed obvious and most natural for centuries only to be superseded by radical shifts.

Treatise on Time, the Beginning and the End date back at least twenty-five centuries. Does the flow of time have an objective, universal meaning beyond human perception? Or, is it fundamentally only a convenient, and perhaps merely psychological, notion? Did the physical universe have a finite beginning or has it been evolving eternally? Leading thinkers across cultures meditated on these issues and arrived at definite but strikingly different answers. For example, in the sixth century BCE, Gautama Buddha taught that "a period of time" is a purely conventional notion, time and space exist only in relation to our experience, and the universe is eternal. In the Christian thought, by contrast, the universe had a finite beginning and there was debate whether time represents "movement" of bodies or if it flows only in the soul. In the fourth century CE, St. Augustine held that time itself started with the world.

[1] Thus I will not include any derivations but instead provide references where the details can be found.

Founding fathers of modern Science from Galileo to Newton continued to accept that God created the universe. Nonetheless, their work led to a radical change of paradigm. Before Newton, boundaries between the absolute and the relative, the true and the apparent and the mathematical and the common were blurry. Newton rescued time from the psychological *and* the material world and made it objective and absolute. It now ran uniformly from the infinite past to the infinite future. This paradigm became a dogma over centuries. Philosophers often used it to argue that the universe itself *had* to be eternal. For, as Immanuel Kant emphasized, otherwise one could ask "what was there before?"

General relativity toppled this Newtonian paradigm in one fell swoop. Now the gravitational field is encoded in space-time geometry. Since geometry is a dynamical, physical entity, it is now perfectly feasible for the universe to have had a finite beginning – the big-bang – at which not only matter but *space-time itself* is born. If space is compact, matter *as well as space-time* end in the big-crunch singularity. In this respect, general relativity took us back to St. Augustine's paradigm but in a detailed, specific and mathematically precise form. In semi-popular articles and radio shows, relativists now like to emphasize that the question "what was there before?" is rendered meaningless because the notions of "before" requires a pre-existing space-time geometry. We now have a new paradigm, a new dogma: In the Beginning there was the Big Bang.

But as I pointed out in Section 7.1, general relativity is incomplete and there is no reason to trust its predictions near space-time singularities. We must fuse it with quantum physics and let the new theory tell us what happens when matter and geometry enter the Planck regime.

7.2.2 Some Key Questions

If the smooth continuum of Minkowski and Einstein is only an approximation, on the issue of the origin of the universe we are now led to ask:

- How close to the big-bang does a smooth space-time of general relativity make sense? Inflationary scenarios, for example, are based on a space-time continuum. Can one show from some first principles that this is a safe approximation already at the onset of inflation?
- Is the big-bang singularity naturally resolved by quantum gravity? This possibility led to the development of the field of quantum cosmology in the late 1960s. The basic idea can be illustrated using an analogy to the theory of the hydrogen atom. In classical electrodynamics the ground state energy of this system is unbounded below. Quantum physics intervenes and, thanks to a non-zero Planck's constant, the ground state energy is lifted to a finite value, $-me^4/2\hbar^2 \approx -13.6$eV. Since it is the Heisenberg uncertainly principle that lies at the heart of this resolution and since the principle must feature also in quantum gravity, one is led to ask: Can a similar mechanism resolve the big-bang and big crunch singularities of general relativity?

- Is a new principle/ boundary condition at the big bang or the big crunch essential? The most well known example of such a boundary condition is the "no boundary proposal" of Hartle and Hawking [16]. Or, do quantum Einstein equations suffice by themselves even at the classical singularities?
- Do quantum dynamical equations remain well-behaved even at these singularities? If so, do they continue to provide a deterministic evolution? The idea that there was a pre-big-bang branch to our universe has been advocated in several approaches, most notably by the pre-big-bang scenario in string theory [17] and ekpyrotic and cyclic models [18, 19] inspired by the brane world ideas. However, these are perturbative treatments which require a smooth continuum in the background. Therefore, their dynamical equations break down at the singularity whence, without additional input, the pre-big-bang branch is not joined to the current post-big-bang branch by a deterministic evolution. Can one improve on this situation?
- If there is a deterministic evolution, what is on the "other side"? Is there just a quantum foam from which the current post-big-bang branch is born, say a "Planck time after the putative big-bang"? Or, was there another classical universe as in the pre-big-bang and cyclic scenarios, joined to ours by deterministic equations?

Clearly, to answer such questions we cannot start by assuming that there is a smooth space-time in the background. But already in the classical theory, it took physicists several decades to truly appreciate the dynamical nature of geometry and to learn to do physics without recourse to a background. In quantum gravity, this issue becomes even more vexing.[2]

For simple systems, including Minkowskian field theories, the Hamiltonian formulation generally serves as the royal road to quantum theory. It was therefore adopted for quantum gravity by Dirac, Bergmann, Wheeler and others. But absence of a background metric implies that the Hamiltonian dynamics is generated by constraints [21]. In the quantum theory, physical states are solutions to quantum constraints. All of physics, including the dynamical content of the theory, has to be extracted from these solutions. But there is no external time to phrase questions about evolution. Therefore we are led to ask:

- Can we extract, from the arguments of the wave function, one variable which can serve as *emergent time* with respect to which the other arguments "evolve"? If not, how does one interpret the framework? What are the physical (i.e., Dirac) observables? In a pioneering work, DeWitt proposed that the determinant of the three-metric can be used as an "internal" time [22]. Consequently, in much of the literature on the Wheeler-DeWitt (WDW) approach to quantum cosmology, the scale factor is assumed to play the role of time, although sometimes only implicitly. However, in closed models the scale factor fails to be monotonic due to classical recollapse and cannot serve as a global time variable already in

[2] There is a significant body of literature on issue; see, e.g., [20] and references therein. These difficulties are now being discussed also in the string theory literature in the context of the AdS/CFT conjecture.

the classical theory. Are there better alternatives at least in the simple setting of quantum cosmology? If not, can we still make physical predictions?

Finally there is an ultraviolet-infrared tension.

- Can one construct a framework that cures the short-distance difficulties faced by the classical theory near singularities, while maintaining an agreement with it at large scales?

By their very construction, perturbative and effective descriptions have no problem with the second requirement. However, physically their implications can not be trusted at the Planck scale and mathematically they generally fail to provide a deterministic evolution across the putative singularity. Since the non-perturbative approaches often start from deeper ideas, it is conceivable that they could lead to new structures at the Planck scale which modify the classical dynamics and resolve the big-bang singularity. But once unleashed, do these new quantum effects naturally "turn-off" sufficiently fast, away from the Planck regime? The universe has had some *14 billion years* to evolve since the putative big bang and even minutest quantum corrections could accumulate over this huge time period leading to observable departures from dynamics predicted by general relativity. Thus, the challenge to quantum gravity theories is to first create huge quantum effects that are capable of overwhelming the extreme gravitational attraction produced by matter densities of some 10^{105} g/cc near the big bang, and then switching them off with extreme rapidity as the matter density falls below this Planck scale. This is a huge burden!

These questions are not new; some of them were posed already in the late sixties by quantum gravity pioneers such as Peter Bergmann, Bryce DeWitt, Charles Misner and John Wheeler [21–23]. However, the field reached an impasse in the late eighties. Fortunately, this status-quo changed significantly over the last decade with a dramatic inflow of new ideas from many directions. In the next two subsections, I will summarize the current status of these issues in loop quantum cosmology.

7.2.3 FRW Models and The WDW Theory

Almost all phenomenological work in cosmology is based on the $\mathrm{k} = 0$ homogeneous and isotropic Friedmann Robertson Walker (FRW) space-times and perturbations thereof. For concreteness, I will focus on FRW model in which the only matter source is a scalar massless field.[3] I will consider $\mathrm{k} = 0$ (or spatially flat) as well as $\mathrm{k} = 1$ (spatially closed) models with or without a cosmological constant (of either sign). Conceptually, these models are interesting for our purpose because *every* of their classical solutions has a singularity (see Fig. 7.1). Therefore a natural singularity resolution without external inputs is highly non-trivial. In light of the

[3] Our discussion will make it clear that it is relatively straightforward to allow additional fields, possibly with complicated potentials.

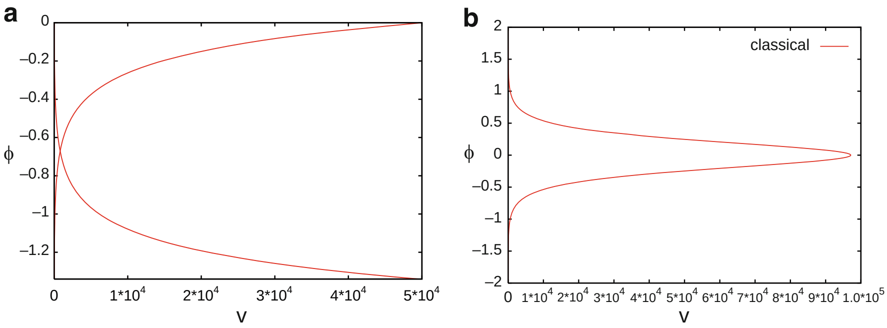


Fig. 7.1 (**a**) Classical solutions in k = 0, $\Lambda = 0$ FRW models with a massless scalar field. Since $p_{(\phi)}$ is a constant of motion, a classical trajectory can be plotted in the v-ϕ plane, where v is the volume (essentially in Planck units). There are two classes of trajectories. In one the universe begins with a big-bang and expands and in the other it contracts into a big crunch. (**b**) Classical solutions in the k = 1, $\Lambda = 0$ FRW model with a massless scalar field. The universe begins with a big bang, expands to a maximum volume and then undergoes a recollapse to a big crunch singularity. Since the volume is double valued in any solution, it cannot serve as a global time coordinate in this case. The scalar field on the other hand does so both in the k = 0 and k = 1 cases

spectacular observational inputs over the past decade, the k = 0 model is the one that is phenomenologically most relevant. However as we will see, because of its classical recollapse, the k = 1 model offers a more stringent viability test for the quantum cosmology.

In the classical theory, one considers one space-time at a time and although the metric of that space-time is dynamical, it enables one to introduce time coordinates that have direct physical significance. However in the quantum theory – and indeed already in the phase space framework that serves as the stepping stone to quantum theory – we have to consider all possible homogeneous, isotropic space-times. In this setting one can introduce a natural foliation of the four-manifold each leaf of which serves as the "home" to a spatially homogeneous three-geometry. However, unlike in non-gravitational theories, there is no preferred physical *time variable* to define evolution. A natural strategy is to use part of the system as an "internal" clock with respect to which the rest of the system evolves. This leads one to Leibnitz's *relational time*. Now, in any spatially homogeneous model with a massless scalar field ϕ, the conjugate momentum $p_{(\phi)}$ is a constant of motion, whence ϕ is monotonic along any dynamical trajectory. Thus, in the classical theory, it serves as a global clock (see Fig. 7.1). Questions about evolution can thus be phrased as: "If the curvature or matter density or an anisotropy parameter is such and such when $\phi = \phi_1$ what is it when $\phi = \phi_2$?" What is the situation in the quantum theory? There is no a priori guarantee that a variable which serves as a viable time parameter in the classical theory will continue to do so in the quantum theory. Whether it does so depends on the form of the Hamiltonian constraint. For instance as Fig. 7.1a shows, in the k = 0 model without a cosmological constant, volume (or the scale factor) is a global clock along any classical trajectory. But the form of the quantum Hamiltonian constraint [24] in loop quantum gravity is such that it does not serve this role in the quantum theory. The scalar field, on the other hand, continues to do so (also in the $k = 1$ case and with or without a cosmological constant).[4]

Because of the assumption of spatial homogeneity, in quantum cosmology one has only a finite number of degrees of freedom. Therefore, although the conceptual problems of quantum gravity remain, there are no field theoretical infinities and one can hope to mimic ordinary text book quantum mechanics to pass to quantum theory.

However, in the $k = 0$ case, because space is infinite, homogeneity implies that the action, the symplectic structure and Hamiltonians all diverge since they are represented as integrals over all of space. Therefore, in any approach to quantum cosmology – irrespective of whether it is based on path integrals or canonical methods – one has to introduce an elementary cell $\mathcal{C}$ and restrict all integrals to it. In actual calculations, it is generally convenient also to introduce a fiducial three-metric ${}^o q_{ab}$ (as well as frames ${}^o e^a_i$ adapted to the spatial isometries) and represent the physical metric q_{ab} via a scale factor a, $\;q_{ab} = a^2\, {}^o q_{ab}$. Then the geometrical dynamical variable can be taken to be either a, or the oriented volume v of the fiducial cell $\mathcal{C}$ as measured by the physical frame e^a_i, where v is positive if e^a_i has the same orientation as ${}^o e^a_i$ and negative if the orientations are opposite. (In either case the physical volume of the cell is $|v|$.) In this chapter I will use v rather than the scale factor. Note, however, physical results cannot depend on the choice of the fiducial $\mathcal{C}$ or ${}^o q_{ab}$.[5] In the $k = 1$ case, since space is compact, a fiducial cell is unnecessary and the dynamical variable v is then just the physical volume of the universe.

With this caveat out of the way, one can proceed with quantization. Situation in the WDW theory can be summarized as follows. This theory emerged in the late sixties and was analyzed extensively over the next decade and a half [21]. Many of the key physical ideas of quantum cosmology were introduced during this period [22, 23] and a number of models were analyzed. However, since a mathematically coherent approach to quantization of full general relativity did not exist, there were no guiding principles for the analysis of these simpler, symmetry reduced systems. Rather, quantization was carried out following "obvious" methods from ordinary quantum mechanics. Thus, in quantum kinematics, states were represented by square integrable wave functions $\Psi(v, \phi)$, where v represents geometry and ϕ, matter; and operators $\hat{v}, \hat{\phi}$ acted by multiplication and their conjugate momenta by ($-i\hbar$ times) differentiation. With these choices The Hamiltonian constraint takes the form of a differential equation that must be satisfied by the physical states [27]:

$$\partial^2_\phi \underline{\Psi}(v, \phi) = \underline{\Theta_o \Psi}(v, \phi) \;:=\; -12\pi G\, (v\partial_v)^2\, \Psi(v, \phi) \tag{7.1}$$

---

[4] If there is no massless scalar field, one could still use a suitable matter field as a "local" internal clock. For instance in the inflationary scenario, because of the presence of the potential the inflaton is not monotonic even along classical trajectories. But it is possible to divide dynamics into "epochs" and use the inflaton as a clock locally, i.e., within each epoch [25]. There is considerable literature on the issue of internal time for model constrained systems [20] (such as a system of two harmonic oscillators where the total energy is constrained to be constant [26]).

[5] This may appear as an obvious requirement but unfortunately it is often overlooked in the literature. The claimed physical results often depend on the choice of $\mathcal{C}$ and/or ${}^o q_{ab}$ although the dependence is often hidden by setting the volume v_o of $\mathcal{C}$ with respect to ${}^o q_{ab}$ to 1 (in unspecified units) in the classical theory.

for k = 0, and

$$\partial_\phi^2 \underline{\Psi}(v,\phi) = -\underline{\Theta}_1 \underline{\Psi}(v,\phi) \;:=\; -\underline{\Theta}_o \underline{\Psi}(v,\phi) - G\, C\, |v|^{\frac{4}{3}}\, \underline{\Psi}(v,\phi)\,, \qquad (7.2)$$

for k = 1, where C is a numerical constant. *In what follows* $\underline{\Theta}$ *will stand for either* $\underline{\Theta}_o$ *or* $\underline{\Theta}_1$. In the older literature, the emphasis was on finding and interpreting the WKB solutions of these equations (see, e.g., [28]). However, near the singularity, the WKB approximation fails and we need an exact quantum theory.

The exact theory can be readily constructed [24, 27]. Note first that the form of (7.1) and (7.2) is the same as that of a Klein-Gordon equation in a two-dimensional static space-time (with a ϕ-independent potential in the k = 1 case), where ϕ plays the role of time and v of space. This suggests that we think of ϕ as the relational time variable with respect to which v, the "true" degree of freedom, evolves. A systematic procedure based on the so-called group averaging method [29] (which is applicable for a very large class of constrained systems) then leads us to the physical inner product between these states. Not surprisingly it coincides with the expression from the Klein-Gordon theory in static space-times.

The physical sector of the final theory can be summarized as follows. The physical Hilbert space $\mathcal{H}_{\rm phy}$ in the k = 0 and k = 1 cases consists of "positive frequency" solutions to (7.1) and (7.2) respectively. A complete set of observables is provided by the momentum $\hat{p_{(\phi)}}$ and the relational observable $\hat{|v|}|_\phi$ representing the volume at the "instant of time ϕ":

$$\hat{p}(\phi) = -i\hbar\partial_\phi \quad \text{and} \quad \hat{V}|_\phi = e^{i\underline{\Theta}(\phi-\phi_o)}\,|v|\,e^{-i\underline{\Theta}(\phi-\phi_o)} \qquad (7.3)$$

There are *Dirac* observables because their action preserves the space of solutions to the constraints and are self-adjoint on the physical Hilbert space $\mathcal{H}_{\rm phy}$. With the exact quantum theory at hand, we can ask if the singularities are naturally resolved. More precisely, from $\hat{p}(\phi)$ and $\hat{V}|_\phi$ we can construct observables corresponding to matter density $\hat{\rho}$ (or space-time scalar curvature $\hat{R}$). Since the singularity is characterized by divergence of these quantities in the classical theory, in the quantum theory we can proceed as follows. We can select a point (v_o, ϕ_o) at a 'late time' ϕ_o on a classical trajectory of Fig. 7.1 – e.g., now, in he history of our universe – when the density and curvature are *very* low compared to the Planck scale, and construct a semi-classical state which is sharply peaked at v_o at $\phi = \phi_o$. We can then evolve this state *backward* in time. Does it follow the classical trajectory? To have the correct "infra-red" behavior, it must, until the density and curvature become very high. What happens in this "ultra-violet" regime? Does the quantum state remain semi-classical and follow the classical trajectory into the big bang? Or, does it spread out making quantum fluctuations so large that although the quantum evolution does not break down, there is no reasonable notion of classical geometry? Or, does it remain peaked on some trajectory which however is so different from the classical one that, in this backward evolution, the the universe "bounces" rather than being crushed into the singularity? Or, does it ... Each of these scenarios provides a distinct pre-

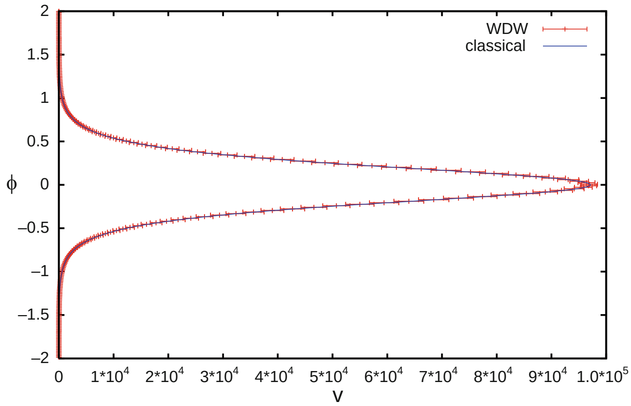


Fig. 7.2 Expectation values (and dispersions) of $|\hat{v}|_\phi$ for the WDW wave function in the $k = 1$ model. The WDW wave function follows the classical trajectory into the big-bang and big-crunch singularities (In this simulation, the parameters were: $p_\phi^\star = 5000$, and $\Delta p_\phi / p_\phi^\star = 0.02$)

diction for the ultra-violet behavior and therefore for physics in the deep Planck regime.[6]

It turns out that the WDW theory leads to similar predictions in both $k = 0$ and $k = 1$ cases [24,27,31]. They pass the infra-red tests with flying colors (see Fig. 7.2). But unfortunately the state follows the classical trajectory into the big bang (and in the $k = 1$ case also the big crunch) singularity. Thus the first of the possibilities listed above is realized. The singularity is not resolved because expectation values of density and curvature continue to diverge in epochs when their classical counterparts do. The analogy to the hydrogen atom discussed in Section 7.2.2 fails to be realized.

7.2.4 Loop Quantum Cosmology: New Quantum Mechanics

For a number of years, the failure of the WDW theory to naturally resolve the big bang singularity was taken to mean that quantum cosmology cannot, by itself, shed any light on the quantum nature of the big bang. Indeed, for systems with a finite number of degrees of freedom we have the von Neumann uniqueness theorem which

[6] Sometimes apparently weaker notions of singularity resolution are discussed. Consider two examples [30]. One may be able to show that the wave function vanishes at points of the classically singular regions of the configuration space. However, if the *physical* inner product is non-local in this configuration space – as the group averaging procedure often implies – such a behavior of the wave function would not imply that the probability of finding the universe at these configurations is zero. The second example is that the wave function may become highly non-classical. This by itself would not mean that the singularity is avoided unless one can show that the expectation values of a family of Dirac observables which become classically singular remain finite there.

guarantees that quantum kinematics is unique. The only freedom we have is in factor ordering and this was deemed insufficient to alter the status-quo provided by the WDW theory.

The situation changed dramatically in LQG. Here, a well established, rigorous kinematical framework *is* available for full general relativity [2–5]. If one mimics it in symmetry reduced models, one is led to a quantum theory which is *inequivalent to that of the WDW theory already at the kinematic level*. Quantum dynamics built in this new arena agrees with the WDW theory in "tame" situations but differs dramatically in the Planck regime, leading to a natural resolution of the big bang singularity.

But what about the von Neumann uniqueness theorem? The theorem states that one-parameter groups $U(\lambda)$ and $V(\mu)$ satisfying the Weyl commutation relations[7] admit (up to isomorphism) a unique irreducible representation by unitary operators on a Hilbert space $\mathcal{H}$ in which $U(\lambda)$ and $V(\mu)$ are weakly continuous in the parameters λ and μ. By Stone's theorem, weak continuity is the necessary and sufficient condition for $\mathcal{H}$ to admit self adjoint operators $\hat{x}, \hat{p}$ such that $U(\lambda) = e^{i\lambda\hat{x}}$ and $V(\mu) = e^{i\mu\hat{p}}$. Therefore assumption of the von Neumann theorem are natural in non-relativistic quantum mechanics and we are led to a unique quantum kinematics. However, in full loop quantum gravity, x is analogous to the gravitational connection and $U(\lambda)$ to its holonomy. One can again construct an abstract algebra using holonomies and operators conjugate to connections and ask for its representations satisfying natural assumptions the most important of which is the diffeomorphism invariance dictated by background independence. There is again a uniqueness theorem [32]. However, in the representation that is thus singled out, holonomy operators – analogs of $U(\lambda)$ – fails to be weakly continuous whence there are no operators corresponding to connections! Furthermore, a number of key features of the theory – such as the emergence of a quantum Riemannian geometry in which there is fundamental discreteness – can be traced back to this unforeseen feature. Therefore, upon symmetry reduction, although we have a finite number of degrees of freedom, it would be incorrect to just mimic Schröddinger quantum mechanics and impose weak continuity. When this assumption is dropped, the von Neumann theorem is no longer applicable and *we have new quantum mechanics* [33].

Thus, the key difference between LQC and the WDW theory lies in the fact that while one does not have reliable quantum kinematics in the WDW theory, there is a well developed and rigorous framework in LQG which, furthermore, is *unique*! If we mimic it as closely as possible in the symmetry reduced theories, we are led to a new kinematic arena, distinct from the one used in the WDW quantum cosmology. LQC is based on this arena.

[7] These are: $U(\lambda V(\lambda) = e^{i\lambda\mu}V(\mu)U(\lambda)$ and can be obtained by setting $U(\lambda) = e^{i\lambda\hat{x}}$ and $V(\mu) = e^{i\mu\hat{p}}$ in the standard Schrödinger theory. Given a representation $U(\lambda)$ is said to be *weakly continuous* in λ if its matrix elements between any two fixed quantum states are continuous in λ.

7.2.5 LQC: Dynamics

It turns out WDW dynamics is not supported by the new arena because, when translated in terms of gravitational connections and their conjugate momenta, it requires that there be an operator corresponding to the connection itself. Therefore one has to develop quantum dynamics ab-initio on the new arena. The result is that the differential operator $\underline{\Theta}_o = -12\pi G\,(v\partial_v)^2$ in Eqs. (7.1) and (7.2) is now replaced by a second order *difference* operator in v, where the step size is dictated by the "area gap" of LQG, i.e., the lowest non-zero eigenvalue of the area operator in LQG. There is a precise sense in which the Wheeler-DeWitt equations result as the limits of LQC equations when the area gap is taken to zero, i.e., when the Planck scale discreteness of quantum geometry determined by LQG is neglected. We will now see that this discreteness is completely negligible at late times but plays a crucial role in the Planck scale geometry near singularities.

The LQC dynamics has been analyzed using three different methods.

- Numerical solutions of the exact quantum equations [24, 27, 31, 34]. A great deal of effort was spent in ensuring that the results are free of artifacts of simulations, do not depend on the details of how semi-classical states are constructed and hold for a wide range of parameters.
- Effective equations [27, 31, 35]. These are differential equations which include the leading quantum corrections. The asymptotic series from which these contributions were picked was constructed rigorously but is based on assumptions whose validity has not been established. Nonetheless the effective equations approximate the exact numerical evolution of semi-classical states extremely well.
- Exactly soluble, but simplified model in the $k = 0$ case [36, 37]. The simplification is well controlled [37]. This analysis has provided some results which provide an analytical understanding of numerical results and also several other results which are not restricted to states which are semi-classical at late times. In this sense the analysis shows that the overall picture is robust within these models.

I will provide a global picture that has emerged from these investigations, first for the $k = 1$ model without the cosmological constant Λ and for the $k = 0$ case for various values of Λ.

Recall that in classical general relativity, the $k = 1$ closed universes start out with a big bang, expand to a maximum volume $V_{\rm max}$ and then recollapse to a big-crunch singularity. Consider a classical solution in which $V_{\rm max}$ is astronomically large – i.e., on which the constant of motion $p_{(\phi)}$ takes a large value $p^\star_{(\phi)}$ – and consider a time ϕ_o at which the volume $v^\star$ of the universe is also large. Then there are well-defined procedures to construct states $\Psi(v, \phi)$ in the *physical Hilbert space* which are sharply peaked at these values of observables $\hat{p}_{(\phi)}$ and $\hat{V}_{\phi_o}$ at the "time" ϕ_o. Thus, at "time" ϕ_o, the quantum universe is well approximated by the classical one. What happens to such quantum states under evolution? As emphasized earlier, there are infra-red and ultra-violet challenges:

1. Does the state remain peaked on the classical trajectory in the low curvature regime? Or, do quantum geometry effects accumulate over the cosmological

time scales, causing noticeable deviations from classical general relativity? In particular, is there a recollapse and if so does the value $V_{\max}$ of maximum volume agree with that predicted by general relativity [38]?

2. What is the behavior of the quantum state in the Planck regime? Is the big-bang singularity resolved? What about the big-crunch? If they are both resolved, what is on the "other side"?

Numerical simulations show that the wave functions do remain sharply peaked on classical trajectories in the low curvature region also in LQC. But there is a radical departure from the WDW results in the strong curvature region. The WDW evolution follows classical dynamics all the way into the big-bang and big crunch singularities (see Fig. 7.2). In LQC, by contrast, *the big bang and the big crunch singularities are resolved and replaced by big-bounces* (see Fig. 7.3). In these calculations, the required notion of semi-classicality turns out to be surprisingly weak: these results hold even for universes with $a_{\max} \approx 23\ell_{\rm Pl}$ and the 'sharply peaked' property improves greatly as $a_{\max}$ grows.

More precisely, numerical solutions have shown that the situation is as follows (for details, see [31]).

- The trajectory defined by the expectation values of the physical observable $\hat{V}|_\phi$ in the full quantum theory is in good agreement with the trajectory defined by the classical Friedmann dynamics until the energy density ρ in the matter field is about 2% of the Planck density. In the classical solution, scalar curvature and the matter energy density keep increasing on further evolution, eventually leading to

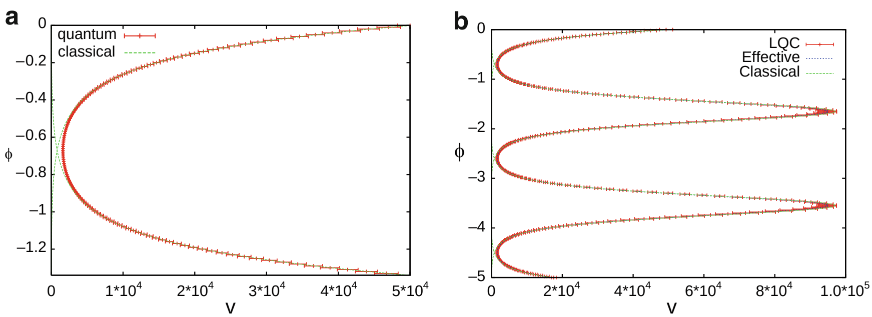


Fig. 7.3 In the LQC evolution of models under consideration, the big bang and big crunch singularities are replaced by quantum bounces. Expectation values and dispersion of $|\hat{v}|_\phi$, are compared with the classical trajectory and the trajectory from effective Friedmann dynamics. The classical trajectory deviates significantly from the quantum evolution at Planck scale and evolves into singularities. The effective trajectory provides an excellent approximation to quantum evolution at all scales. (**a**) The $k=0$ case. In the backward evolution, the quantum evolution follows our post big-bang branch at low densities and curvatures but undergoes a quantum bounce at matter density $\rho \sim 0.82\rho_{\rm PL}$ and joins on to the classical trajectory that was contracting to the future. (**b**) The $k=1$ case. The quantum bounce occurs again at $\rho \sim 0.82\rho_{\rm Pl}$. Since the big bang and the big crunch singularities are resolved the evolution undergoes cycles. In this simulation $p^\star_{(\phi)} = 5\times 10^3$, $\Delta p_{(\phi)}/p^\star_{(\phi)} = 0.018$, and $v^\star = 5\times 10^4$

a big bang (respectively, big crunch) singularity in the backward (respectively, forward) evolution, where $v \to 0$. The situation is very different with quantum evolution. As the density and curvature increases further, quantum geometry effects become dominant creating an effective repulsive force which rises very quickly, overwhelms classical gravitational attraction, and causes a bounce at $\rho \sim 0.82\rho_{\rm Pl}$, thereby resolving the past (or the big bang) and future (or the big crunch) singularities. There is thus a cyclic scenario depicted in Fig. 7.3.

- The volume of the universe takes its minimum value $V_{\rm min}$ at the bounce point. $V_{\rm min}$ scales linearly with $p_{(\phi)}$:[8]

$$V_{\rm min} = \left(\frac{4\pi G\gamma^2\Delta}{3}\right)^{\frac{1}{2}}\ p_{(\phi)}\ \approx (1.28\times 10^{-33}\,{\rm cm})\ p_{(\phi)} \tag{7.4}$$

Consequently, $V_{\rm min}$ can be *much* larger than the Planck size. Consider for example a quantum state describing a universe which attains a maximum radius of a megaparsec. Then the quantum bounce occurs when the volume reaches the value $V_{\rm min} \approx 5.7 \times 10^{16}\,{\rm cm}^3$, *some* 10^{115} *times the Planck volume*. Deviations from the classical behavior are triggered when the density or curvature reaches the Planck scale. The volume can be very large and is not the relevant scale for quantum gravity effects.

- After the quantum bounce the energy density of the universe decreases and the repulsive force dies quickly when matter density reduces to about two percept of the Planck density. The quantum evolution is then well-approximated by the classical trajectory. On subsequent evolution, the universe recollapses both in classical and quantum theory at the value $V = V_{\rm max}$ when energy density reaches a minimum value $\rho_{\rm min}$. $V_{\rm max}$ scales as the 3/2-power of $p_{(\phi)}$:

$$V_{\rm max} = (16\pi G/3\ell_o^2)^{3/4}\ p_{(\phi)}^{3/2}\ \approx\ 0.6\, p_{(\phi)}^{3/2} \tag{7.5}$$

Quantum corrections to the classical Friedmann formula $\rho_{\rm min} = 3/8\pi G a_{\rm max}^2$ are of the order $O(\ell_{\rm Pl}/a_{\rm max})^4$. For a universe with $a_{\rm max} = 23\ell_{\rm Pl}$, the correction is only one part in 10^5. For universes which grow to macroscopic sizes, classical general relativity is essentially exact near the recollapse.

- Using ideas from geometrical quantum mechanics [39], one can obtain certain effective classical equations which incorporate the leading quantum corrections [31, 35]. While the classical Friedmann equation is $(\dot a/a)^2 = (8\pi G/3)\,(\rho - 3/8\pi G a^2)$, the effective equation turns out to be

$$\left(\frac{\dot a}{a}\right)^2 = \frac{8\pi G}{3}\,(\rho - \rho_1(v))\,\left[f(v) - \frac{\rho}{\rho_{\rm crit}}\right] \tag{7.6}$$

[8] Here and in what follows, numerical values are given in the classical units $G = c = 1$. In these units $p_{(\phi)}$ has the same physical dimensions as $\hbar$ and the numerical value of $\hbar$ is $2.5 \times 10^{-66}{\rm cm}^2$.

where ρ_1 and f are specific functions of v with $\rho_1 \sim 3/8\pi G a^2$. Bounces occur when $\dot{a}$ vanishes, i.e., at the value of v at which the matter density equals $\rho_1(v)$ or $f(v) = \rho/\rho_{\rm crit}$. The first root $\rho(v) = \rho_1(v)$ corresponds to the classical recollapse while the second root, $f(v) = \rho/\rho_{\rm crit}$, to the quantum bounce. Away from the Planck regime, $f \approx 1$ and $\rho/\rho_{\rm crit} \approx 0$. Bounces occur when $\dot{a}$ vanishes, i.e., at the value of v at which the matter density equals $\rho_1(v)$ or $\rho_2(v)$.

- For quantum states under discussion, the density $\rho_{\rm max}$ is well approximated by $\rho_{\rm crit} \approx 0.82\rho_{\rm Pl}$ up to terms $O(\ell_{\rm Pl}^2/a_{\rm min}^2)$, independently of the details of the state and values of $p_{(\phi)}$. (For a universe with maximum radius of a megaparsec, $\ell_{\rm Pl}^2/a_{\rm min}^2 \approx 10^{-76}$.) The density $\rho_{\rm min}$ at the recollapse point also agrees with the value $(3/8\pi G a_{\rm max}^2)$ predicted by the classical evolution to terms of the order $O(\ell_{\rm Pl}^4/a_{\rm max}^4)$. Furthermore the scale factor $a_{\rm max}$ at which recollapse occurs in the quantum theory agrees to a very good precision with the one predicted by the classical dynamics.
- The trajectory obtained from effective Friedmann dynamics is in excellent agreement with quantum dynamics *throughout the evolution.* In particular, the maximum and the minimum energy densities predicted by the effective description agree with the corresponding expectation values of the density operator $\hat{\rho} \equiv \widehat{p_{(\phi)}^2/|p|^3}$ computed numerically.
- The state remains sharply peaked for a *very large number of "cycles"*. Consider the example of a semi-classical state with an almost equal relative dispersion in $p_{(\phi)}$ and $|v|_\phi$ and peaked at a large classical universe of the size of a megaparsec. When evolved, it remains sharply peaked with relative dispersion in $|v|_\phi$ of the order of 10^{-6} *even after* 10^{50} *cycles of contraction and expansion!* Any given quantum state eventually ceases to be sharply peaked in $|v|_\phi$ (although it continues to be sharply peaked in the constant of motion $p_{(\phi)}$). Nonetheless, the quantum evolution continues to be deterministic and well-defined for an infinite number of cycles. This is in sharp contrast with the classical theory where the equations break down at singularities and there is no deterministic evolution from one cycle to the next.

This concludes the summary of our discussion of the k = 1 model. An analogous detailed analysis has been carried out also in the k = 0 model, again with a free massless scalar field [24, 27, 34, 37]. In this case, if the cosmological constant Λ vanishes, as Fig. 7.1 shows, classical solutions are of two types, those which start out at the big-bang and expand out to infinity and those which start out with large volume and contract to the big crunch singularity. Again, in this case while the WDW solution follows the classical trajectories into singularities, the LQC solutions exhibit a big bounce. The LQC dynamics is again faithfully reproduced by an effective equation: the Friedmann equation $(\dot{a}/a)^2 = (8\pi G\,\rho/3)$ is replaced just by $(\dot{a}/a)^2 = (8\pi G\,\rho/3)\,(1-\rho/\rho_{\rm crit})$. The quantum correction $\rho/\rho_{\rm crit}$ is completely negligible even at the onset of the standard inflationary era. Quantum bounce occurs at $\rho = \rho_{\rm crit}$ and the critical density is again given by $\rho_{\rm crit} \approx 0.82\rho_{\rm Pl}$. Furthermore, one can show that the spectrum of the density operator *on the physical Hilbert space* admits a finite upper bound $\rho_{\rm sup}$. By plugging values of constants in the analytical expression of this bound, one finds $\rho_{\rm sup} = \rho_{\rm crit}$! If $\Lambda > 0$, there are again two types of classical trajectories but the one which starts out at the big-bang expands to an infinite volume in finite value $\phi_{\rm max}$ of ϕ. (The other trajectory is a "time reverse" of this.) Because the ϕ "evolution" is unitary in LQC, it yields a natural extension of the classical solution beyond $\phi_{\rm max}$. If $\Lambda < 0$, the classical universe undergoes a recollapse. This is faithfully reproduced by the LQC evolution. Since both the big-bang and the big-crunch singularities are resolved, the LQC evolution leads to a cyclic universe as in the k = 1 model. Thus, in all these cases, the principal features of the LQC evolution are robust, including the value of $\rho_{\rm crit}$.

Let us summarize the overall situation. In simple cosmological models, all the questions raised in Section 7.2.2 have been answered in LQC in remarkable detail. The scalar field plays the role of an internal or emergent time and enables us to interpret the Hamiltonian constraint as an evolution equation. The matter momentum $\hat{p}(\phi)$ and "instantaneous" volumes $\hat{V}|_\phi$ form a complete set of Dirac observables and enable us to ask physically interesting questions. Answers to these questions imply that the big bang and the big crunch singularities are naturally replaced by quantum bounces. On the "other side" of the bounce there is again a large universe. General relativity is an excellent approximation to quantum dynamics once the matter density falls below a couple of percent of the Planck density. Thus, LQC successfully meets both the ultra-violet and infra-red challenges. Furthermore results obtained in a number of models using distinct methods re-enforce one another. One is therefore led to take at least the qualitative findings seriously: *Big bang is not the Beginning nor the big crunch the End.* Quantum space-time appears to be vastly larger than what general relativity had us believe!

7.3 Black Holes

The idea of black holes is quite old. Already in 1784, in an article in the Proceedings of the Royal Society John Mitchell used the formula for escape velocity in Newtonian gravity to argue that light can not escape from a body of mass M if it is compressed to a radius $R = 2GM/c^2$. He went on to say

> if there should exist in nature any [such] bodies .... we could have no information from sight; yet if any other luminous bodies should happen to revolve around them we might still perhaps from the motions of these revolving bodies infer the existence of the central ones with some degree of probability.

Remarkably, it is precisely observations of this type that have now led us to the conclusion that there is a 3.4 million solar mass black hole in the center of our galaxy! In the second volume of Exposition du systèm du Monde published in 1798, the Marquis de Laplace came to the same conclusion independently and was more confident of the existence of black holes:

> there exist, in the immensity of space, opaque bodies as considerable in magnitude, and perhaps equally as numerous as stars.

While in many ways these observations are astonishingly prescient, the underlying reasoning is in fact incorrect. For, if light (which is assumed to be corpuscular in this argument) from a distant source were to impinge on such an object, it would bounce back and by Newtonian conservation laws it would reach the point from which it came. Distant observers should therefore be able to see these objects. Indeed, if all speeds – including that of light – are relative as in Newtonian mechanics, there can really be no black holes. The existence of black holes requires both gravity and an absolute speed of light; general relativity is essential.

7.3.1 Horizons

To capture the intuitive notion that black hole is a region from which signals can not escape to the asymptotic part of space-time, one needs a precise definition of future infinity. The standard strategy is to use Penrose's conformal boundary $\mathcal{I}^+$ [40]. It is a future boundary: No point of the physical space-time lies to the future of any point of $\mathcal{I}^+$. It has topology $\mathbb{S}^2 \times \mathbb{R}$ and it is null (assuming that the cosmological constant is zero). In Minkowski space-time, one can think of $\mathcal{I}^+$ as the "final resting place" of all future directed null geodesics. More precisely, the chronological past $I^-(\mathcal{I}^+)$ of $\mathcal{I}^+$ is entire Minkowski space.[9]

Given a general asymptotically flat space-time (M, g_{ab}), one first finds the chronological past $I^-(\mathcal{I}^+)$ of $\mathcal{I}^+$. If it is not the entire space-time, then there is a region in (M, g_{ab}) from which one cannot send causal signals to infinity. When this happens, one says that the space-time admits a black hole. More precisely, *Black-hole region* $\mathbb{B}$ of (M, g_{ab}) is defined as

$$\mathbb{B} = M \; - \; I^-(\mathcal{I}^+) \tag{7.7}$$

where the right side is the set of points of M which are not in $I^-(\mathcal{I}^+)$. The boundary $\partial\mathbb{B}$ of the black hole region is called the *event horizon* (EH) and is denoted by E [41]. $I^-(\mathcal{I}^+)$ is often referred to as the asymptotic region and e is the boundary of this region within physical space-time.

Event horizons and their properties have provided a precise arena to describe black holes and their dynamics. In particular, we have the celebrated result of Hawking's [41, 42]: assuming energy conditions, the area $a_{\rm hor}$ of an EH cannot decrease under time evolution. The area $a_{\rm hor}$ is thus analogous to thermodynamic entropy. There are other laws governing black holes which are in equilibrium (i.e., stationary) and that make transitions to nearby equilibrium states due to influx of energy and angular momentum. They are similar to the zeroth and the first law of

[9] $I^-(\mathcal{I}^+)$ is the set of all points in the physical space-time from which there is a future directed time-like curve to a point on $\mathcal{I}^+$ in the conformally completed space-time. The term "chronological" refers to the use of time-like curves. A curve which is everywhere time-like or null is called "causal".

thermodynamics and suggest that the surface gravity κ of stationary black holes is the analog of thermodynamic temperature. These analogies were made quantitative and precise by an even deeper result Hawking obtained using quantum field theory in a black hole background [43]: black holes radiate quantum mechanically as though they are black bodies at temperature $T = \kappa\hbar/2\pi$. Their entropy is then given by $S = a_{\rm hor}/4\ell_{\rm Pl}^2$. Not surprisingly these results have led to a rich set of insights and challenges over the last 35 years.

However, the notion of an EH also has two severe limitations. First, while the notion neatly captures the idea that asymptotic observers can not "look into" a black hole, it is too global for many applications. For example, since it refers to null infinity, it can not be used in spatially compact space-times. Asymptotic flatness and the notion of $\mathcal{I}^+$ is used also in other contexts, in particular to discuss gravitational radiation in full, non-linear general relativity [40]. However, there $\mathcal{I}^+$ is used just to facilitate the imposition of boundary condition and make notions such as "$1/r^n$-fall-off" precise. Situation with EHs is quite different because they refer to the *full chronological past of* $\mathcal{I}^+$. As a consequence, by changing the geometry in a small – say Planck scale region – around the singularity, one can change the EH dramatically and even make it disappear [1]! As I explained in Section 7.1, there is no reason to trust classical general relativity very close to the singularity. If the singularity is resolved due to quantum effects, there may be no longer an EH. What then is a black hole? If the notion continues to be meaningful, can we still associate with it entropy in absence of an EHs?

The second limitation is that the notion is teleological; it lets us speak of a black hole *only after we have constructed the entire space-time*. Thus, for example, an EH may well be developing in the room you are now sitting *in anticipation* of a gravitational collapse that may occur in this region of our galaxy a million years from now. Indeed, as Fig. 7.4a shows, EHs can form and grow *even in flat space-time* where there is no influx of matter of radiation. How can we then attribute direct physical significance to the growth of their area? Clearly, when astrophysicists say that they have discovered a black hole in the center of our galaxy, they are referring to something much more concrete and quasi-local than an EH.

Over the last 5 years, quasi-local horizons were introduced to improve on this situation [44–47]. The idea is to use the notion of marginally trapped surfaces. Consider a space-like 2-sphere in Minkowski space and illuminate it instantaneously. Then there are two light fronts, one traveling outside the sphere and expanding continuously and the other traveling inside and contracting. Now, if the 2-sphere were placed in a strong gravitational field, both these light fronts could contract. Then light would be trapped and the sphere would not be visible from outside. These two situations are separated by the marginal case where one light front would be contracting and the area of the other would neither decrease not increase. Such 2-surfaces are said to be *marginally trapped* and their world tubes represent quasi-local horizons. More precisely, a marginally trapped tube (MTT) is a 3-manifold which is foliated by a family of marginally trapped 2-spheres. If it is space-like, the area of the marginally trapped surfaces increases to the future and the MTT is called a *dynamical horizon* (DH). Heuristically it represents a growing black hole.

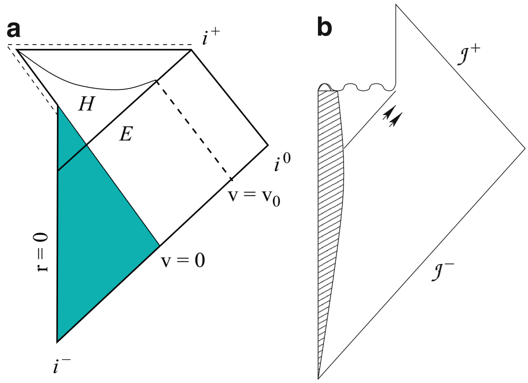


Fig. 7.4 (**a**) A Vaidya solution: Collapse of a spherical null fluid to form a black hole. The null fluid radiation starts at the retarded time $v = 0$ and ends at $v = v_0$. Space time is flat in the past of $v = 0$ and Schwarzschild to the future of $v = v_0$. The dynamical horizon H starts out space-like and joins on to the null event horizon at $v = v_0$. The event horizon first forms and grows in the flat part of space-time. (**b**) Conjectured Penrose diagram of an evaporating black hole: A black hole forms by stellar collapse and evaporates due to Hawking radiation. Due to back reaction, the singularity loses its strength as we move right along the wiggly line, and finally disappears. Nonetheless because there is still a piece of space-like singularity in the future, $\mathcal{I}^+$ does not constitute the full future boundary of space-time, leading to information loss

If the MTT is null, it is called an *isolated horizon* (IH) and represents a black hole in equilibrium. In Fig. 7.4a a DH H forms due to gravitational collapse of infalling null fluid, grows in area with the in-fall and settles down to an IH which coincides with the future part of the EH E. Note that the definitions of MTT, DH and IH are all quasi-local. In particular, they are not teleological; you can be rest assured that none of these quasi-local horizons exists in the room you are now sitting in!

There is however a significant drawback: lack of uniqueness. Although partial uniqueness results exist [48], in general we cannot yet associate a unique DH with a generic, growing black hole. But this weakness is compensated in large measure by the fact that interesting results hold for *every* DH. In particular, not only does the direct analog of Hawking's area theorem hold on DHs, but there is a precise *quantitative* relation between the growth of area of a DH and the amount of energy falling into it [45, 46]. Therefore, in striking contrast with EHs, we can associate a direct physical significance to the growth in area of DHs. This and other quantitative relations have already made DHs very useful in numerical simulations of black hole formation and mergers [47]. Finally, since they refer only to the space-time geometry in their immediate vicinity, the existence and properties of these horizons are insensitive to what happens near the singularity. Thus, quantum gravity modifications in the space-time geometry in the vicinity of the classical singularity would have no effect on these horizons.[10]

Conceptually these quasi-local horizons are also useful in quantum considerations. Let us first consider equilibrium situations. In loop quantum gravity, there is a statistical mechanical derivation of the entropy associated with any isolated horizon [2, 49]. These cover not only the familiar stationary black holes but also hairy black holes as well as cosmological horizons. Next, consider dynamics. During the collapse, the MTT is space-like and we have a DH. But once the in-fall of matter ends, the mass of the black hole must decrease and the horizon area must shrink. In this phase the MTT is time-like and so there is no obstruction at all for leakage of matter from inside the MTT to the outside region.

To summarize, black holes were first described using EHs. While this description has led to important insights, they also have some important limitations in the dynamical context. The more recent quasi-local horizons provide concepts and tools that are more directly useful both in numerical relativity and quantum gravity.

7.3.2 Hawking Radiation and Information Loss

Consider a spherically symmetric gravitational collapse depicted in Fig. 7.4a. Once the black hole is formed, space-time develops a new, future boundary at the singularity, whence one can not reconstruct the geometry and matter fields by evolving the data *backward* from future null infinity, $\mathcal{I}^+$. Thus, whereas an appropriately chosen family of observers near $\mathcal{I}^-$ has full information needed to construct the entire space-time, no family of observers near $\mathcal{I}^+$ has such complete information. In this sense, black hole formation leads to information loss. Note that, contrary to the heuristics often invoked, this phenomenon is not directly related to black hole uniqueness results: it occurs even when uniqueness theorems fail, as with "hairy" black holes or in presence of matter rings non-trivially distorting the horizon. The essential ingredient is the future singularity which can act as the sink of information.

A natural question then is: what happens in quantum gravity? Is there again a similar information loss? Hawking's [43] celebrated work of 1974, mentioned in Section 7.3.1, analyzed this issue in the framework of quantum field theory in curved space-times. In this approximation, three main assumptions are made: (i) the gravitational field can be treated classically; (ii) one can neglect the back-reaction of the spontaneously created matter on the space-time geometry; and (iii) the matter quantum field under investigation is distinct from the collapsing matter, so one can focus just on spontaneous emission. Under these assumptions, at late times there is

[10] One might wonder: Don't the singularity theorems essentially guarantee that if there is an MTT there must be a singularity? Recall however that the theorems also assume classical Einstein's equations and certain energy conditions. Both these assumptions would be violated in quantum gravity. Therefore, it is perfectly feasible for MTTs to exist even though the (quantum) space-time has no singularities.

a steady emission of particles to $\mathcal{I}^+$ and the spectrum is thermal at a temperature dictated by the surface gravity of the final black hole. In particular, pure states on $\mathcal{I}^-$ evolve to mixed states on $\mathcal{I}^+$. However, this external field approximation is too crude; in particular it violates energy conservation. To cure this drawback, one can include back-reaction. A detailed calculation is still not available. However, following Hawking [43], one argues that, as long as the black hole is large compared to the Planck scale, the quasi-stationary approximation should be valid. Then, by appealing to energy conservation and the known relation between the mass and the horizon area of *stationary* black holes, one concludes that the area of the EH should steadily decrease.[11] This then leads to black hole evaporation depicted in Fig. 7.4b [42]. If one does not examine space-time geometry but uses instead intuition derived from Minkowskian physics, one may be surprised that although there is no black hole at the end, the initial pure state has evolved in to a mixed state. Note however that even after the inclusion of back reaction, in this scenario *there is still a final singularity, i.e., a final boundary in addition to* $\mathcal{I}^+$. Therefore, it is not at all surprising that, in this approximation, information is lost – it is still swallowed by the final singularity. Thus, provided Fig. 7.4b is a reasonable approximation of black hole evaporation and one does not add new input "by hand", then pure states must evolve in to mixed states.

The question then is to what extent this diagram is a good representation of the physical situation. The general argument in the relativity community has been the following. Figure 7.4b should be an excellent representation of the actual physical situation as long as the black hole is much larger than the Planck scale. Therefore, problems, if any, are associated *only* with the end point of the evaporation process. It is only here that the semi-classical approximation fails and one needs full quantum gravity. Whatever these 'end effects' are, they deal only with the Planck scale objects and would be too small to recover the correlations that have been steadily lost as the large black hole evaporated down to the Planck scale. Hence pure states must evolve to mixed states and information is lost.

Tight as this argument seems, it overlooks two important considerations. First, one would hope that quantum theory is free of infinities whence Fig. 7.4b can not be a good depiction of physics near the *entire singularity* – not just near the end point of the evaporation process. Second, as we saw in Section 7.3.1, the EH is a highly global and teleological construct. Since the structure of the *quantum* space-time could be very different from that of Fig. 7.4b near (and "beyond") the singularity, the causal relations implied by the presence of the EH of Fig. 7.4b is likely to be quite misleading [1]. Indeed, using the AdS/CFT conjecture, string theorists have argued that the evolution must be unitary and information is not lost. However, since the crux of that argument is based on the boundary theory (which is conjectured to be equivalent to string theory in the bulk), this line of reasoning does not provide a direct *space-time description* of how and why the information is recovered. Where does the above reasoning of relativists, fail? How must it be corrected?

[11] This does not contradict the area law because the energy conditions used in its derivation are violated by the quantum emission.

I believe that answer to these question lies in the fact that, because of singularity resolution, the quantum space-time is larger than the classical [50]. In support of this view, in the next two sections I will use a two-dimensional black hole to argue that the loss of information is not inevitable even in space-time descriptions favored by relativists.

7.3.3 CGHS Black Holes

Let us begin with the spherical collapse of a massless scalar field f in four space-time dimensions resulting in a black hole. Because of spherical symmetry, it is convenient to factor out by the 2-spheres of symmetry and pass to the $r - t$ plane. Let us express the four-dimensional space-time metric ${}^4g_{ab}$ as:

$$ {}^4g_{ab} = g_{ab} + \frac{e^{-2\phi}}{\kappa^2}\, s_{ab} \,, $$

where we have introduced a constant κ with dimensions of inverse length and set $r = e^{-\phi}/\kappa^2$. Then the (symmetry reduced) Einstein Hilbert action becomes

$$ \begin{aligned} &S(g, \phi, \kappa) \\ &= \frac{1}{2G} \int d^2x\, \sqrt{|g|}\, \big[e^{-2\phi}\, (R + \mathbf{2}\nabla^a\phi\nabla_a\phi + \mathbf{2e}^{-2\phi}\kappa^2)\ + G\, \mathbf{e}^{-\phi}\nabla^a f \nabla_a f\big] \end{aligned} \tag{7.8} $$

where R is the scalar curvature of the 2-metric g. This theory is very rich especially because the well-known critical phenomena. The classical equations cannot be solved exactly. However, an apparently small modification of this action – indicated by bold faced terms – gives a two-dimensional theory which *is* exactly soluble classically. There is again a black hole formed by gravitational collapse and it evaporates by Hawking radiation. This is the Callen, Giddings, Harvey, Strominger (CGHS) model [51] and it arose upon symmetry reduction of a low energy action motivated by string theory. Because it has many of the qualitative features of the 4-d theory but is technically simpler, the model attracted a great deal of attention in the 1990s (for reviews, see, e.g., [52]). Here, the basic fields are again a two-dimensional metric g of signature -+, a geometrical scalar field ϕ, called the dilaton, and a massless scalar field f. The action is given by:

$$ S(g, \phi, f) := \frac{1}{2G} \int d^2x\, \sqrt{|g|}\, \big[e^{-2\phi}\, (R + \mathbf{4}\nabla^a\phi\nabla_a\phi + \mathbf{4}\kappa^2) + G\, \nabla^a f \nabla_a f\big] \tag{7.9} $$

We will analyze this 2-d theory in its own right.

Recall that imposition of spherical symmetry in 4-d general relativity implies that the gravitational field is completely determined by matter – the true degrees of freedom are all contained in the matter. The same is true in the CGHS model. Furthermore, in the CGHS case there is a simplification: the equation of motion of f is just $\Box_{(g)} f = 0$; dynamics of f is decoupled from ϕ. In two dimensions, the physical metric g^{ab} is conformally related to a flat metric $g^{ab} = \Omega \eta^{ab}$ and conformal invariance of the wave equation implies that $\Box_{(g)} f = 0$ if and only if $\Box_{(\eta)} f = 0$. Therefore, we can fix a fiducial flat metric η, parameterize g by Ω and determine f by solving the wave equation on the two-dimensional Minkowski space (M_o, η). Finally, let us set $\Phi = e^{-2\phi}$ and write the conformal factor Ω as $\Omega = \Theta^{-1} \Phi$. The passage $(g, \phi, f) \longrightarrow (\Theta, \Phi, f)$ just corresponds to a convenient choice of field redefinitions.

Since $\Box_{(\eta)} f = 0$, we know $f = f_+(z^+) + f_-(z^-)$, where $f_\pm$ are arbitrary smooth functions of their arguments and $z^\pm$ are the advanced and retarded coordinates of η (i.e., $\eta_{ab} = -\partial_{(a} z^+ \, \partial_{b)} z^-$). Given f, the equations of motion for Θ and Φ (together with appropriate boundary conditions) determine the classical solution completely. To display it, it is simplest to use coordinates $x^\pm$ given by:

$$\kappa x^+ = e^{\kappa z^+} \quad \text{and} \quad \kappa x^- = -e^{-\kappa z^-} .$$

Then, for any given $f_\pm$, the solution is given by

$$\Theta = -\kappa^2 x^+ x^- \quad \text{and}$$
$$\Phi = \Theta - \frac{G}{2} \int_0^{x^+} \mathrm{d}\bar{x}^+ \int_0^{\bar{x}^+} \mathrm{d}\bar{\bar{x}}^+ \, (\partial f_+ / \partial \bar{\bar{x}}^+)^2 - \frac{G}{2} \int_0^{x^-} \mathrm{d}\bar{x}^- \int_0^{\bar{x}^-} \mathrm{d}\bar{\bar{x}}^- \, (\partial f_- / \partial \bar{\bar{x}}^-)^2 . \tag{7.10}$$

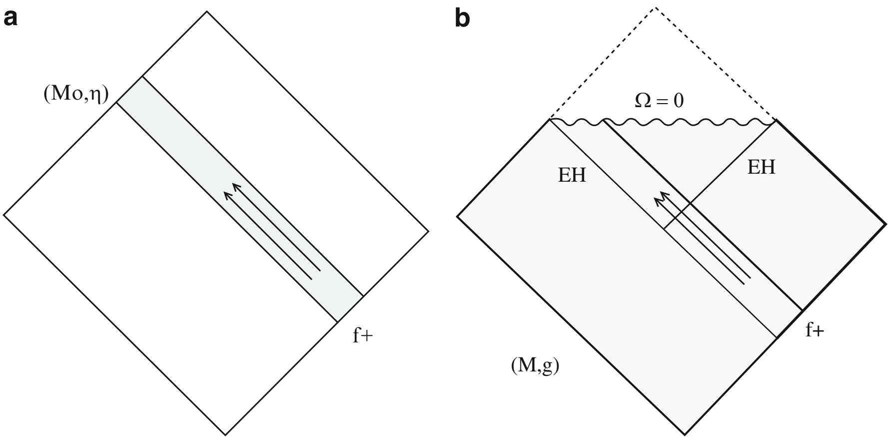


Fig. 7.5 (**a**) A typical solution for the f_+ mode in Minkowski space. (**b**) When interpreted in terms of the physical metric g, a black hole has formed because of the gravitational collapse of f_+. The physical space-time M is a proper subset of M_o but the subset realized depends on the solution f_+. Therefore already in the classical Hamiltonian theory, the kinematical arena is provided by M_o

The black hole sector of interest is obtained by setting $f_- = 0$ as in Fig. 7.5a and letting f_+ collapse. (Alternatively, one could $f_+ = 0$ and consider the collapse of f_-.)

But why is there a black hole? Fields f_+, Θ, Φ are all smooth on the entire manifold M_o. Recall, however, that the physical metric is given by $g^{ab} = \Omega \eta^{ab} \equiv \Theta^{-1} \Phi\, \eta^{ab}$. On the entire manifold M_o, Θ is smooth and nowhere vanishing. However, it is easy to verify that Φ vanishes along a space-like line (see Fig. 7.5b). On this line g^{ab} becomes degenerate and its scalar curvature diverges. Thus there is a space-like singularity; the physical space-time manifold M on which g_{ab} is well defined is only a part of the fiducial Minkowski manifold M_o (see Fig. 7.5b). Is it hidden behind an event horizon? To ask this question, we should first verify that (M, g_{ab}) admits a *complete* [53] future null infinity $\mathcal{I}^+$ and the past of $\mathcal{I}^+$ does not contain the singularity. In two space-times dimensions, past as well as future null infinity has two pieces, one to the right and the other to the left and they are joined only by points $i^\pm$ at time-like infinity. In the solutions under consideration $\mathcal{I}^+_{\rm R}$ is complete but $\mathcal{I}^+_{\rm L}$ is not. Therefore strictly we can meaningfully ask if there is a black hole only with respect to $\mathcal{I}^+_{\rm R}$ and the answer is in the affirmative. Fortunately to analyze the Hawking radiation and information loss, we can focus just on $\mathcal{I}^+_{\rm R}$. Before going on to these issues, it is interesting to note that there is a black hole in spite of the fact that the solution (f, θ, Φ) is perfectly regular. This is because the physical meaning of the solution has to be analyzed using the physical geometry determined by g.

Solution (7.10) represents a black hole formed by the gravitational collapse of f_+. In the spirit of Hawking's original derivation, let us study the dynamics of a *test* quantum field $\hat{f}_-$ on this black hole geometry. Now $\mathcal{I}^-$ of every physical metric g coincides with the past null infinity $\mathcal{I}^{o-}$ of Minkowski space (M_o, η) and $g = \eta$ in a neighborhood of $\mathcal{I}^{o-}_{\rm L}$. So we can begin with the vacuum state $|0\rangle_-$ at $\mathcal{I}^{o-}_{\rm L}$ and ask for its dynamical content. In the Heisenberg picture, the operators evolve and state remain fixed. The issue then is that of interpretation of the fixed state $|0\rangle_-$ in the geometry given by g in a neighborhood of $\mathcal{I}^+_{\rm R}$. Now, two important factors of the geometry come into play. First, although the physical metric g is asymptotically flat, *it does not agree with η even at $\mathcal{I}^+_{\rm R}$*. More precisely, the affine parameter y^- at $\mathcal{I}^+_{\rm R}$ is a non-trivial function of z^-, reflecting the fact that the asymptotic time translation of g does not coincide with any of the asymptotic time translations of η. Therefore there is a mixing of positive and negative frequency modes. Since $|0\rangle_-$ is defined using z^-, it is populated with particles defined at $\mathcal{I}^+_{\rm R}$ by g. Second, $\mathcal{I}^+_{\rm R}$ is a proper subset of $\mathcal{I}^{o+}_{\rm R}$. Therefore, we have to trace over modes of $\hat{f}_-$ with support on $\mathcal{I}^{o+}_{\rm R} - \mathcal{I}^+_{\rm R}$. Therefore, as far as measurements of observables near $\mathcal{I}^+_{\rm R}$ are concerned, the state $|0\rangle_-$ is indistinguishable from a density matrix ρ on the Hilbert space $\mathcal{H}$ of $\hat{f}_-$ at $\mathcal{I}^+_{\rm R}$. Detailed calculation shows that at late times, ρ is precisely the thermal state at temperature $\hbar\kappa/2\pi$ [54]! Thus, in the CGHS model there is indeed Hawking radiation and therefore, by repeating the reasoning summarized in Section 7.3.2 one can conclude that there must be information loss.

I will conclude this section by summarizing the similarities and differences in the four and two dimensional analyses. In both cases there is a formation of a black hole due to gravitational collapse and the test quantum field is distinct from the field that collapses. Thanks to asymptotic flatness at past null infinity, the vacuum state $|0\rangle_-$ of the test field is well-defined and the key issue is that of its physical interpretation in the physical geometry near future null infinity. Finally although κ was introduced as a constant in the CGHS theory, one can verify that it is in fact the surface gravity of the stationary black hole in the future of the support of f_+. In both cases the Hawking temperature is this given by $\hbar/2\pi$ times the surface gravity. However, there are also some important differences. First, whereas there is just one $\mathcal{I}^-$ and $\mathcal{I}^+$ in four dimensions in the CGHS case we have two copies of each and the clear-cut black hole interpretation holds only with respect to $\mathcal{I}^+_{\rm R}$. Second, while in four dimensions κ and hence the Hawking temperature is inversely proportional to the mass of the black hole, in the CGHS case it is a constant. Finally, at a technical level, even in the spherically symmetric reduction of the four dimensional theory, the equation satisfied by the scalar field f is much more complicated than the CGHS wave equation. Therefore, while analysis of the CGHS black hole does provide valuable insights for the four dimensional case, one cannot take directly over results.

7.3.4 Quantum Geometry

Since the model is integrable classically, many steps in the passage to quantum theory are simplified [15]. Our basic fields will again be $\hat{f}, \hat{\Theta}, \hat{\Phi}$. The true degree of freedom is in the scalar field f and it satisfies just the wave equation on Minkowski space (M_o, η). Therefore, it is straightforward to construct the Fock space $\mathcal{F} = \mathcal{F}_+ \otimes \mathcal{F}_-$ and represent $\hat{f}_\pm$ as operator valued distributions on $\mathcal{F}$. Classically, we have explicit expressions (7.10) of fields Θ and Φ in terms of f on all of M_o. In quantum theory, because of trace anomaly the equations satisfied by $\hat{\Theta}, \hat{\Phi}$ are more complicated. Therefore explicit solutions are not available. However, these are hyperbolic equations on the fiducial Minkowski space and the boundary values *at* $\mathcal{I}^{o-}$ are given by the (unambiguous) operator versions of (7.10). Therefore, in principle, it should be possible to solve them. A conjecture based on approximate solutions is that $\hat{\Theta}$ would be an operator field and $\hat{\Phi}$ an operator valued distribution on $\mathcal{F}$.

At first, may seem surprising that there is no Hilbert space corresponding to geometry. However, already at the classical level the covariant phase space can be coordinatized completely by the scalar field f and geometric fields Θ, Φ are just functionals on this phase space. The situation in quantum theory is precisely what one would expect upon quantization. While the full quantum theory is still incomplete in the CGHS model, there is a simpler and interesting system in which this feature is realized in detail: cylindrical gravitational waves in four-dimensional general relativity. This system is equivalent to $2+1$ Einstein gravity coupled to an axi-symmetric scalar field. Again because there are no gravitational degrees of freedom in $2+1$ dimensions, the true degree of freedom can be encoded in the scalar field which now satisfies a wave equation in a fiducial $2+1$ dimensional Minkowski space. The regulated metric operator is represented as an operator valued distribution on the Fock space of the scalar field [55] and leads to interesting and unforeseen quantum effects [56].

Returning to the CGHS model, we can now ask: What is a quantum black hole? In the classical theory, black holes result if we specify a smooth profile f^o as initial data for f_+ on $\mathcal{I}_{\rm R}^{o-}$ and zero data for f_- on $\mathcal{I}_{\rm L}^{o-}$. In quantum theory, then, a candidate black hole would a quantum state $|\Psi\rangle$ which is peaked at this classical data on $\mathcal{I}^{o-}$: $|\Psi\rangle = |0\rangle_- \otimes |C_f^o\rangle_+$ where $|0\rangle_-$ is the vacuum state in $\mathcal{F}_-$ and $|C_f^o\rangle_+$ is the coherent state in $\mathcal{F}_+$ peaked at the classical profile f_o of f_+. One can show that if one solves the quantum equations for $\hat{\Theta}, \hat{\Phi}$ in a certain approximation (the first step in a certain bootstrapping), then the states Ψ do emerge as black holes: the expectation values of $\hat{g}^{ab}, \hat{\Phi}$ are precisely those of the classical black hole solutions. In particular, $\langle\hat{\Phi}\rangle$ vanishes along a space-like line which appears as the singularity in the classical theory. However, the true *quantum* geometry near this classical singularity is perfectly regular [15]: the *operator* $\hat{\Phi}$ does not vanish, only its expectation value does. Furthermore, one can also calculate fluctuations and show that they are small near infinity but huge near the classical singularity. Consequently, the expectation values are poor representations of the actual quantum geometry in a neighborhood of the classical singularity. The fact that the quantum metric $\hat{g}^{ab}$ is regular on M_o already in this approximation suggests that the singularity may be resolved in the quantum theory making the quantum space-time larger than the classical one. There is then a possibility that there may be no information loss.

This issue is probed using the mean field approximation (MFA) [15]. Here, one first takes the expectation value of the the quantum equations governing $\hat{\Theta}, \hat{\phi}$ in the state Ψ and, furthermore, replaces Φ, θ by their expectation values. Thus, for example, $\langle\hat{\Theta}\,\hat{\Phi}\rangle$ is replaced by $\langle\hat{\Theta}\rangle\ \langle\hat{\Phi}\rangle$ but $\langle :(\partial\hat{f})^2 :\rangle$ is kept as is. This amounts to ignoring the quantum fluctuations in the geometric operators $\hat{\Theta}, \hat{\Phi}$ but not those in the matter field $\hat{f}$. This approximation can be justified in the limit in which there is a large number N of scalar fields $\hat{f}$ rather than just one and we restrict ourselves to regions in which "fluctuations in the geometry are less than N times the fluctuations in any one matter field". In this region, the mean field approximation provides a good representation of the geometry that includes back reaction of the Hawking radiation.

It turns out that the resulting equations on $\bar{\Theta} := \langle\hat{\Theta}\rangle$ and $\bar{\Phi} := \langle\hat{\Phi}\rangle$ were already obtained sometime ago using functional integral techniques and solved numerically [57]. By making appeal to the four-dimensional theory whose symmetry reduction gives the CGHS models, one can introduce the notion of marginally trapped surfaces and their area. Simulations showed that marginally trapped surfaces do form due to infalling matter, the marginally trapped tube is first space-like – i.e., is a dynamical horizon – but, after the inflow of collapsing matter ends, becomes time-like due to the leakage of the Hawking radiation. Thus the scenario based on quasi-local horizons is realized. In the dynamical horizon phase, the horizon area $a_{\rm hor}$ increases and in the subsequent Hawking evaporation, it decreases to zero: It is again the MTT that evaporates. However, further evolution to the future moves one closer to what was the classical singularity. As I mentioned above, in this region the quantum fluctuations in geometry become huge and so the mean field approximation

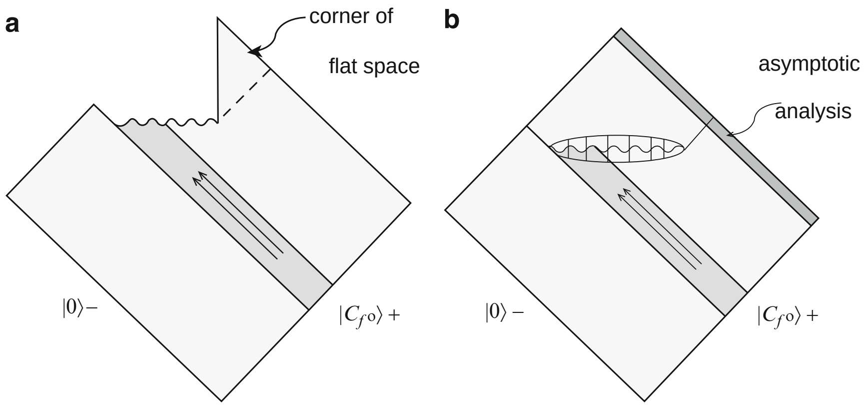


Fig. 7.6 (**a**) The CGHS analog of the Penrose diagram 7.4b. This diagram has been used for a number of years to describe space-time geometry after inclusion of back reaction. Singularity is still part of the future boundary and so the information is lost. (**b**) The space-time diagram suggested by the asymptotic analysis of mean field equations near $\mathcal{I}_R^+$. In the quantum space-time $\mathcal{I}_R^+$ is 'as long as' $\mathcal{I}_R^{o+}$, whence $|0\rangle_-$ is a pure state also with respect to the physical metric g. It is however populated by particles and resembles thermal density matrix at an intermediate region of $\mathcal{I}_R^+$

fails. The simulations cannot be continued further. However, since the area of the marginally trapped surface shrunk to zero, it was assumed – as is reasonable – that the Bondi mass at the corresponding retarded instant of time would be zero on $\mathcal{I}_R^+$. Therefore, following what Hawking did in four-dimensions, it became customary to attach by hand a corner of Minkowski space to the numerically evolved space-time thereby arriving at a Penrose diagram of Fig. 7.6a. Note that in this diagram, the future boundary for the $\hat{f}_-$ modes consists not just of $\mathcal{I}_R^+$ but also a piece of the singularity. As I argued in Section 7.3.2, if this is an accurate depiction of the physical situation, one would conclude that $|0\rangle_-$ at $\mathcal{I}_L^{o-}$ would evolve to a density matrix on $\mathcal{I}_R^+$ and information would indeed be lost.

Note however that the key to the information loss issue lies in the geometry near future infinity and MFA should be valid there. Thus, rather than attaching a corner of flat space by hand at the end of the numerical simulation, we can use the mean field equations near $\mathcal{I}_R^+$ and let them tell us what the structure of $\mathcal{I}_R^+$ of the *physical* metric is.

To realize this idea, one has to make three assumptions: (i) exact quantum equations can be solved and the expectation value $\bar{g}^{ab}$ of $\hat{g}^{ab}$ admits a smooth right null infinity $\mathcal{I}_R^+$ which coincides with $\mathcal{I}_R^{o+}$ *in the distant past* (i.e. near i_R^o); (ii) MFA holds in a neighborhood of $\mathcal{I}_R^+$; and, (iii) Flux of quantum radiation vanishes at some finite value of the affine parameter y^- of $\mathcal{I}_R^+$ defined by the asymptotic time translation of $\bar{g}$. All three assumptions have been made routinely in the analysis of the information loss issue, although they are often only implicit. Indeed, one cannot even meaningfully ask if information is lost unless the first two hold. (The third assumption can be weakened to allow the flux to decay sufficiently fast in the future.) Then, a systematic analysis of the MFA equations shows [15] that *the right future null infinity $\mathcal{I}_{\mathrm{R}}^{+}$ of the physical metric $\bar{g}$ coincides with that of η*; $\mathcal{I}_{\mathrm{R}}^{+} = \mathcal{I}_{\mathrm{R}}^{o+}$ (see Fig. 7.6). This implies that to interpret $|0\rangle_-$ at $\mathcal{I}_{\mathrm{R}}^{+}$ we no longer have to trace over any modes; in contrast to the situation encountered in the external field approximation discussed in Section 7.3.3, all modes of $\hat{f}_-$ are now accessible to the asymptotically stationary observers of $\bar{g}$. The vacuum state $|0\rangle_-$ of η is pure also with respect to $\bar{g}$. But is it in the asymptotic Fock space of $\bar{g}$? Calculation of Bogoluibov coefficients shows [15] that the answer is in the affirmative. Thus, the interpretation of $|0\rangle_-$ with respect to $\bar{g}$ is that it is a pure state populated by pairs of particles at $\mathcal{I}_{\mathrm{R}}^{+}$. *There is neither information loss nor remnants.*

Let us summarize the discussion of CGHS black holes. A key simplification in this model is that the matter field satisfies just the wave equation on (M_o, η^{ab}). Therefore, given initial data on $\mathcal{I}^{o-}$, we already know the state everywhere both in the classical and the quantum theory. However, the state derives its physical interpretation from geometry which is a complicated functional of the matter field. We do not yet know the quantum geometry everywhere. But approximation methods suggest that $\hat{g}^{ab}$ is likely to be well-defined (and nowhere vanishing) everywhere on M_o. By making rather weak assumptions on the asymptotic behavior of its expectation value $\bar{g}^{ab}$, one can conclude that the right future null infinity $\mathcal{I}_{\mathrm{R}}^{+}$ of $\bar{g}^{ab}$ coincides with $\mathcal{I}_{\mathrm{R}}^{o+}$ of η^{ab} and the affine parameters y^- and z^- defined by the two metrics are such that the exact quantum state $|0\rangle_-$ is a pure state in the asymptotic Fock space of $\bar{g}^{ab}$. The S-matrix is unitary and there is no information loss. Thus the asymptotic analysis leads us to a Penrose diagram of Fig. 7.6b which is significantly different from Fig. 7.6a, based on Hawking's original proposal [43]. In particular, the quantum space-time does not end at a future singularity and is larger than that in Fig. 7.6a. The singularity is replaced by a genuinely quantum region in which quantum fluctuations are large and the notion of a smooth metric tensor field is completely inadequate. However, in contrast to the situation in quantum cosmology of Section 7.2, a full solution to the quantum equations is still lacking.

7.4 Discussion

In Section 7.2 we saw that many of the long standing questions regarding the big bang have been answered in detail in the FRW cosmologies with a massless scalar field and the results are physically appealing. Main departures from the WDW theory occur due to *quantum geometry effects* of LQG. There is no fine tuning of initial conditions, nor a boundary condition at the singularity, postulated from outside. Also, there is no violation of energy conditions. Indeed, quantum corrections to the matter Hamiltonian do not play any role in the resolution of singularities of these models. The standard singularity theorems are evaded because the geometrical side of the classical Einstein's equations is modified by the quantum geometry corrections of LQC. While the detailed results presented in Section 7.2.5 are valid only for these simplest models, partial results have been obtained also in more complicated models indicating that the singularity resolution is rather robust.

In this respect there is a curious similarity with the very discovery of physical singularities in general relativity. They were first encountered in special examples. But the examples were also the physically most interesting ones – e.g., the big-bang and the Schwarzschild curvature singularities. At first it was thought that these space-times are singular because they are highly symmetric. It was widely believed that generic solutions of Einstein'e equations should be non-singular. As is well-known, this belief was shattered by the Penrose-Hawking singularity theorems. Some 40 years later we have come to see that the big bang and the Schwarzschild singularities are in fact resolved by quantum geometry effects. Is this an artifact of high symmetry? Or, are there robust *singularity resolution theorems* lurking just around the corner?

A qualitative picture that emerges is that the non-perturbative quantum geometry corrections are *"repulsive"*.[12] While they are negligible under normal conditions, they dominate when curvature approaches the Planck scale and can halt the collapse that would classically have lead to a singularity. In this respect, there is a curious similarity with the situation in the stellar collapse where a new repulsive force comes into play when the core approaches a critical density, halting further collapse and leading to stable white dwarfs and neutron stars. This force, with its origin in the Fermi-Dirac statistics, is *associated with the quantum nature of matter*. However, if the total mass of the star is larger than, say, five solar masses, classical gravity overwhelms this force. The suggestion from LQC is that a new repulsive force *associated with the quantum nature of geometry* comes into play and is strong enough to counter the classical, gravitational attraction, irrespective of how large the mass is. It is this force that prevents the formation of singularities. Since it is negligible until one enters the Planck regime, predictions of classical relativity on the formation of trapped surfaces, dynamical and isolated horizons would still hold. But assumptions of the standard singularity theorems would be violated. There would be no singularities, no abrupt end to space-time where physics stops. Non-perturbative, background independent quantum physics would continue.

One can also analyze the CGHS models using LQG [58]. However, I used the more familiar Fock spaces to illustrate the fact that the basic phenomenon of singularity resolution by quantum geometry effects is more general. In the CGHS case the analysis is not as complete as in the cosmological models because the CGHS model has an infinite number of degrees of freedom. But results obtained using various approximations strongly suggest that, as in the cosmological case, quantum space-times are larger than what the classical theory suggests. However, nature of the quantum space-time is quite different in the two cases. In the cosmological case the state remains sharply peaked around a smooth geometry even near the bounce.

[12] We saw in Section 7.2.4 that there is no connection operator in LQG. As a result the curvature operator has to be expressed in terms of holonomies and becomes non-local. The repulsive force can be traced back to this non-locality. Heuristically, the polymer excitations of geometry do not like to be packed too densely; if brought too close, they repel.

The expression of the effective metric which provides an excellent approximation to the exact quantum state does have an explicit dependence on $\hbar$ due to quantum corrections. However, it is smooth everywhere. In the CGHS model on the other hand quantum fluctuations in the metric become large in the Planck regime whence one cannot approximate the quantum geometry by *any* smooth geometry. Rather, there is a genuine quantum bridge joining the smooth metric in the distant past to that in the distant future.

At first one might think that, since quantum gravity effects concern only a tiny region, whatever quantum effects there may be, their influence on the global properties of space-time should be negligible whence they would have almost no bearing on the issue of the Beginning and the End. However, as we saw, once the singularity is resolved, vast new regions appear on the 'other side' ushering in new possibilities that were totally unforeseen in the realm of Minkowski and Einstein. Which of them are realized generically? Is there a manageable classification? If, as in the CGHS case, there are domains in which geometry is truly quantum, classical causality would be rendered inadequate to understand the global structure of space-time. Is there a well-defined but genuinely quantum notion of causality which reduces to the familiar one on quantum states which are sharply peaked on a classical geometry? Or, do we just abandon the idea that space-time geometry dictates causality and formulate physics primarily in relational terms? There is a plethora of such exciting challenges. Their scope is vast, they force us to introduce novel concepts and they lead us to unforeseen territories. These just the type of omens that foretell the arrival of a major paradigm shift to take us beyond the space-time continuum of Minkowski and Einstein.

Acknowledgements Much of this chapter is based on joint work with Alex Corichi, Tomasz Pawlowski, Param Singh, Victor Taveras, Madhavan Varadarajan and Kevin Vandersloot. Innumerable discussions with them sharpened my understanding. I have also benefited greatly from comments, suggestions and probing questions of many colleagues, especially Martin Bojowald, James Hartle, Gary Horowitz, Jerzy Lewandowski, Donald Marolf, Roger Penrose and Carlo Rovelli. This work was supported in part by the NSF grants PHY04-56913 and PHY05-54771, the Alexander von Humboldt Foundation, the The George A. and Margaret M. Downsbrough Endowment and the Eberly research funds of Penn State.

References

1. Hajicek, P.: Origin of black hole radiation. Phys. Rev. **D36**, 1065 (1987)
2. Ashtekar, A., Lewandowski, J.: Background independent quantum gravity: A status report Class. Quant. Grav. **21**, R53–R152 (2004)
3. Rovelli, C.: Quantum gravity. Cambridge University Press, Cambridge (2004)
4. Thiemann, T.: Introduction to modern canonical quantum general relativity. Cambridge University Press, Cambridge, (2007)
5. Ashtekar, A., Lewandowski, J., Marolf, D., Mourão, J., Thiemann, T.: Quantization of diffeomorphism invariant theories of connections with local degrees of freedom. Jour. Math. Phys. **36**, 6456–6493 (1995)
6. Rovelli, R., Smolin, L.: Discreteness of area and volume in quantum gravity. Nucl. Phys. **B442**, 593–622 (1995); Erratum **B456**, 753 (1996)

7. Ashtekar, A., Lewandowski, J.: Quantum theory of geometry I: area operators. Class. Quantum Grav. **14**, A55–A81 (1997)
8. Ashtekar, A., Lewandowski, J.: Quantum theory of geometry II: volume operators. Adv. Theo. & Math. Phys. **1**, 388–429 (1997)
9. Perez, A.: Spin foam models for quantum. Class. Quant. Grav. **20**, R43–R104 (2003)
10. Bojowald, M.: Absence of singularity in loop quantum cosmology. Phys. Rev. Lett. **86**, 5227–5230 (2001) arXiv:gr-qc/0102069
11. Bojowald, M.: Loop quantum cosmology. Liv. Rev. Rel. **8**, 11 (2005)
12. Ashtekar, A.: An introduction to loop quantum gravity through cosmology. Nuovo Cimento **112B**, 1–20 (2007) arXiv:gr-qc/0702030
13. Ashtekar, A., Bojowald, M.: Quantum geometry and the Schwarzschild singularity. Class. Quant. Grav. **23**, 391–411 (2006) Modesto, L. Black hole interior from loop quantum gravity, gr-qc/0611043
14. Boehmer, C.G., Vandersloot, K.: Loop quantum dynamics of the schwarzschild interior. Phys. Rev. **D76**, 104030 (2007)
15. Ashtekar, A., Taveras, V., Varadarajan, M.: Information is not lost in 2-dimensional black hole evaporation. arXiv:0801.1811
16. Hartle, J.B., Hawking, S.W.: Wave function of the universe. Phys. Rev. D **28**, 2960 (1983)
17. Gasperini, M., Veneziano, G.: The pre-big bang scenario in string cosmology. Phys. Rept. **373**, 1 (2003) arXiv:hep-th/0207130
18. Khoury, J., Ovrut, B.A., Steinhardt, P.J., Turok, N.: The ekpyrotic universe: colliding branes and the origin of the hot big bang. Phys. Rev. **D64**, 123522 (2001) hep-th/0103239
19. Khoury, J., Ovrut, B., Seiberg, N., Steinhardt, P.J., Turok, N.: From big crunch to big bang. Phys. Rev. **D65**, 086007 (2002) hep-th/0108187
20. Ashtekar, A., Stachel, J. (eds.): Conceptual problems of quantum gravity. Birkhäuser, Boston (1988)
21. Komar, A.: Quantization program for general relativity. In: Carmeli, M., Fickler, S.I., Witten, L. (eds.) Relativity. Plenum, New York (1970); Kuchař, K.: Canonical methods of quantization. In: Isham, C.J., Penrose, R., Sciama, D.W. (eds.) Quantum gravity 2, a second Oxford symposium. Clarendon Press, Oxford (1981)
22. DeWitt, B.S.: Quantum theory of gravity I. The canonical theory. Phys. Rev. **160**, 1113–1148 (1967)
23. Misner, C.W.: Mixmaster universe. Phys. Rev. Lett. **22**, 1071–1074 (1969); Minisuperspace. In: Magic without Magic: John Archibald Wheeler; a collection of essays in honor of his sixtieth birthday. W. H. Freeman, San Francisco (1972)
24. Ashtekar, A., Pawlowski, T., Singh, P.: Quantum nature of the big bang: an analytical and numerical investigation I. Phys. Rev. **D73**, 124038 (2006)
25. Ashtekar, A., Pawlowski, T., Singh, P.: Loop quantum cosmology in the pre-inflationary epoch (in preparation)
26. Rovelli, C.: Quantum mechanics without time: a model. Phys. Rev. **D42**, 2638 (1990)
27. Ashtekar, A., Pawlowski, T., Singh, P.: Quantum nature of the big bang: improved dynamics. Phys. Rev. **D74**, 084003 (2006)
28. Kiefer, C.: Wavepsckets in in minisuperspace. Phys. Rev. **D38**, 1761–1772 (1988)
29. Marolf, D.: Refined algebraic quantization: systems with a single constraint. arXive:gr-qc/9508015; Quantum observables and recollapsing dynamics. Class. Quant. Grav. **12**, 1199–1220 (1994); Ashtekar, A., Bombelli, L., Corichi, A.: Semiclassical states for constrained systems. Phys. Rev. **D72**, 025008 (2005)
30. Kamenshchik, A., Kiefer, C., Sandhofer, B.: Quantum cosmology with big break singularity. Phys. Rev. **D76**, 064032 (2007)
31. Ashtekar, A., Pawlowski, T., Singh, P., Vandersloot, K.: Loop quantum cosmology of $k = 1$ FRW models. Phys. Rev. **D75**, 0240035 (2006); Szulc, L., Kaminski, W., Lewandowski, J.: Closed FRW model in loop quantum cosmology. Class. Quant. Grav. **24**, 2621–2635 (2006)
32. Lewandowski, J., Okolow, A., Sahlmann, H., Thiemann, T.: Uniqueness of diffeomorphism invariant states on holonomy flux algebras. Comm. Math. Phys. **267**, 703–733 (2006); Fleishchack, C.: Representations of the Weyl algebra in quantum geometry. arXiv:math-ph/0407006

33. Ashtekar, A., Bojowald, M. Lewandowski, J.: Mathematical structure of loop quantum cosmology. Adv. Theo. Math. Phys. **7**, 233–268 (2003)
34. Ashtekar, A., Pawlowski, T., Singh, P.: Quantum nature of the big bang. Phys. Rev. Lett. **96**, 141301 (2006), `arXiv:gr-qc/0602086`
35. Willis, J.: On the low energy ramifications and a mathematical extension of loop quantum gravity. Ph.D. Dissertation, The Pennsylvaina State University (2004); Ashtekar, A., Bojowald, M., Willis, J.: Corrections to Friedmann equations induced by quantum geometry, IGPG preprint (2004); Taveras, V.: LQC corrections to the Friedmann equations for a universe with a free scalar field, IGC preprint (2007)
36. Bojowald, M.: Dynamical coherent states and physical solutions of quantum cosmological bounces. Phys. Rev. **D 75**, 123512 (2007)
37. Ashtekar, A., Corichi, A., Singh, P.: Robustness of predictions of loop quantum cosmology. PRD (in press), arXiv:0710.3565
38. Green, D., Unruh, W.: Difficulties with recollapsing models in closed isotropic loop quantum cosmology. Phys. Rev. **D70**, 103502 (2004) `arXiv:gr-qc/04-0074`
39. Ashtekar, A., Schilling, T.A.: Geometrical formulation of quantum mechanics. In: Harvey, A. (ed.) On Einstein's path: essays in honor of Engelbert Schücking. pp. 23–65. Springer, New York (1999) arXiv:gr-qc/9706069
40. Penrose, R.: Zero rest mass fields including gravitation. Proc. R. Soc. (London) **284**, 159–203 (1965)
41. Hawking, S.W.: The event horizon. In: DeWitt, B.S., DeWitt, C.M. (eds.). Black holes, North-Holland, Amsterdam (1972)
42. Hawking, S.W.: Gravitational radiation from colliding black holes. Phys. Rev. Lett. **26**, 1344–1346 (1971)
43. Hawking, S.W.: particle creation by black holes, Commun. Math. Phys. **43**, 199–220 (1975)
44. Ashtekar, A., Beetle, C., Dreyer, O., Fairhurst, S., Krishnan, B., Lewandowski, J., Wisniewski, J.: Generic isolated horizons and their applications. Phys. Rev. Lett. **85**, 3564–3567 (2000)
45. Ashtekar, A., Krishnan, B.: Dynamical horizons: energy, angular momentum, fluxes and balance laws. Phys. Rev. Lett. **89**, 261101–261104 (2002)
46. Ashtekar, A., Krishnan, B.: Dynamical horizons and their properties. Phys. Rev. **D68**, 104030–104055 (2003)
47. Ashtekar, A., Krishnan, B.: Isolated and dynamical horizons and their applications. Living Rev. Rel. **10**, 1–78 (2004) `gr-qc/0407042`
48. Ashtekar, A., Galloway, G.: Some uniqueness results for dynamical horizons. Adv. Theor. Math. Phys. **9**, 1–30 (2005)
49. Ashtekar, A., Baez, J., Corichi, A., Krasnov, K.: Quantum geometry and black hole entropy. Phys. Rev. Lett. **80**, 904–907 (1998); Ashtekar, A., Baez, J., Krasnov, K.: Quantum geometry of isolated horizons and black hole entropy. Adv. Theor. Math. Phys. **4**, 1–94 (2000); Ashtekar, A., Engle, J., Van Den Broeck, C.: Quantum horizons and black hole entropy: inclusion of distortion and rotation. Class. Quant. Grav. **22**, L27–L33 (2005)
50. Ashtekar, A., Bojowald, M.: Black hole evaporation: a paradigm. Class. Quant. Grav. **22**, 3349–3362 (2005)
51. Callen, C.G., Giddings, S.B., Harvey, J.A., Strominger, A.: Evanescent black holes. Phys. Rev. **D45**, R1005–R1010 (1992)
52. Giddings, S.B.: Quantum mechanics of black holes. `arXiv:hep-th/9412138`; Strominger, A.: Les Houches lectures on black holes. `arXiv:hep-th/9501071`
53. Geroch, R., Horowitz, G.: Asymptotically simple does not imply asymptotically Minkowskian. Phys. Rev. Lett. **40**, 203–207 (1978); Erratum: Phys. Rev. Lett. **40**, 483 (1978)
54. Giddings, S.B., Nelson, W.M.: Quantum emission from two-dimensional black holes. Phys. Rev. **D46**, 2486–2496 (1992)
55. Ashtekar, A., Pierri, M.: Probing quantum gravity through exactly soluble midi-superspaces I. J. Math. Phys. **37**, 6250–6270 (1996)
56. Ashtekar, A.: Large quantum gravity effects: Unexpected limitations of the classical theory. Phys. Rev. Lett. **77**, 4864–4867 (1996)

57. Low, D.A.: Semiclassical approach to black hole evaporation. Phys. Rev. **D47**, 2446–2453 (1993); Pirqan, T., Strominger, A.: Numerical analysis of black hole evaporation. Phys. Rev. **D48**, 4729–4734
58. Laddha, A.: Polymer quantization of the CGHS model I. **24**, 4969–4988 (2007); Polymer quantization of the CGHS model II. Class. Quantum Grav. **24**, 4989–5009 (2007)

Chapter 8
Space-Time Extensions in Quantum Gravity

Martin Bojowald

Abstract Space-time as described by special and general relativity is a classical object. In general relativity, it becomes dynamical, participates in the interactions between all degrees of freedom and can even cease to exist when a singularity is reached. As a dynamical object, space-time should be subject to quantization just like matter degrees of freedom, which would change the basic conceptual understanding and also, in general, its dynamics. Corresponding deviations from the classical theory, welcome or unwelcome, are important to test the viability of proposed quantum theories of gravity, but also to understand the fundamental meaning of space-time. Two of the main examples are the singularity problem, where one would like quantum gravity to change the classical dynamics strongly enough to avoid curvature divergences, and the local Lorentz symmetry, which one would like quantum gravity to change not too much. Several recent results for these two cases are discussed based on a loop quantization. Singularities are resolved in cosmological models as a dynamical consequence of discrete quantum space-time which can turn attractive gravity repulsive. The state of the universe existed before the big bang, although it could have been very different from what it appears to be now. Local Lorentz symmetries can be tested in an effective description, where restrictions on quantization ambiguities are already realized. Both features shed light on the nature of quantum space-time.

Keywords Singularities · Big bang · Black holes · Lorentz invariance · Causality

8.1 Introduction

Since Minkowski's promotion of space-time to one single physical entity, it has undergone a series of further transformations. First, with general relativity space-time became a dynamical object subject to its own equations of motion, with equal

M. Bojowald (✉)
Institute for Gravitation and the Cosmos, The Pennsylvania State University, 104 Davey Lab, University Park, PA 16802, USA
e-mail: bojowald@gravity.psu.edu

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_8,

rights to the fields describing matter. It is no longer the rigid stage of Minkowski's, which still determines its local structure, but unfolds in concert with the matter it contains. Disturbingly, space-time as a dynamical object may even come to an end in the abyss of singularities. To resolve this major problem, among other things, combinations of general relativity with quantum physics are being considered. At this stage, conceptual properties of space-time fully merge with those of matter: it is subject to quantum fluctuations and may possess an elementary atomic structure of discrete building blocks.

Now, as a result of this transformative process, the blissful union of space and time is sometimes being challenged: an atomic space-time could, much like a material crystal, act as a dispersive medium for waves and imply Lorentz violating effects such as super-luminal motion even in vacuum. Minkowski space as the very starting point of this long chain of constructions would then be relegated to a mere low energy mirage, a useful tool for calculations but barred from playing a fundamental role in a much more violent quantum theory of gravity at higher energies.

Constructions and calculations in quantum gravity are notoriously difficult to perform, and so the jury on the status of space-time is still out. Here, we will first discuss how quantum gravity can serve to extend the life-span of space-time by removing classical singularities. In this context, the atomic nature of space does play an important role, implying corrections to classical equations of motion such that solutions remain free of singularities. Intuitively, a discrete space-time can store only a finite amount of energy because wave functions of matter fields with too small wavelengths cannot be supported. A consistent dynamical formulation of a discrete space-time must thus provide mechanisms to prevent energy densities from reaching too large values as they would at the classical big bang. Loop quantum cosmology [1] realizes this explicitly by turning the gravitational force repulsive at high energies which can happen in several different ways [2]. Once a universe becomes too small, its collapse must stop and turn into re-expansion. Details will be given in what follows which show that, generically, quantum properties of space-time do have to be taken into account: while classical singularities may become traversable, large quantum fluctuations of space-time can easily arise. Although a given state evolves deterministically through the classical singularity, the quantum nature of space-time beyond this border can vary widely, depending sensitively on the initial state [3, 4].

Quantum fluctuations and higher moments of a state now change the space-time picture because they present additional non-classical degrees of freedom which usually couple to the classical ones. This coupling can be hard to control in a general, inhomogeneous space-time. But there are other implications of a quantum space-time, related to its atomic structure, which are simpler and can be used as a proxy to test the status of Lorentz symmetries in quantum gravity. While any quantum correction could easily destroy classical symmetries, they are subject to consistency conditions. Such conditions can be evaluated effectively in some cases, where they are seen to restore crucial properties of a Lorentz-invariant space-time [5].

First, isotropic cosmology will be considered to illustrate how quantum aspects may change the space-time picture. Thanks to the availability of a solvable model

[6], this situation is under good control, allowing even dynamical coherent states to be analyzed. In the same section, some (weaker) results about black holes will also be considered. In a final section before the conclusions, new features of inhomogeneous space-times and the role of local symmetries will be discussed.

8.2 Space-Time in Classical Cosmology

The simplest way to make Minkowski space with line element $\mathrm{d}s^2 = -\mathrm{d}\tau^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$ dynamical is to allow for a varying length scaled by the scale factor $a(\tau)$: the Friedmann–Robertson–Walker line element $\mathrm{d}s^2 = -\mathrm{d}\tau^2 + a(\tau)^2(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2)$. This is a solution to Einstein's equation when the Friedmann equation

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\frac{H_{\mathrm{matter}}(a)}{a^3} = \frac{8\pi G}{3}\rho \tag{8.1}$$

and the Raychaudhuri equation

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3a^3}\left(H_{\mathrm{matter}}(a,\phi,p_\phi) - a\frac{\partial H_{\mathrm{matter}}(a,\phi,p_\phi)}{\partial a}\right) = -\frac{4\pi G}{3}(\rho+3P) \tag{8.2}$$

with a given matter Hamiltonian H_{matter} are satisfied for $a(\tau)$. (The contribution $\rho = H_{\mathrm{matter}}(a)/a^3$ is the energy density provided by matter, while $P = -\frac{1}{3}a^{-2}\partial H_{\mathrm{matter}}/\partial a = -\partial H_{\mathrm{matter}}/\partial V$ is pressure defined as the negative derivative of energy by volume.) In the vacuum case, we obtain the Minkowski solution $a = \mathrm{const}$, but a dynamical space-time results if matter is present.

For instance, for a free, massless scalar field (a stiff fluid whose pressure equals its energy density) we have $H_{\mathrm{matter}} = \frac{1}{2}a^{-3}p_\phi^2$ where the momentum p_ϕ of the scalar is a constant of motion. Thus, we can easily determine the solution

$$a(\tau)^3 = 2\sqrt{6\pi G}\,p_\phi(\tau-\tau_0) + a_0^3 \tag{8.3}$$

where a_0 is the initial value of a at τ_0. As follows immediately, any solution for the scale factor vanishes at some finite time τ_1, at which point the Friedmann equation no longer presents a well-defined initial value problem. Physically, energy density and curvature diverge at this big bang singularity.

8.3 Atomic Quantum Space-Time

If quantum gravity provides an atomic picture of space and time, just as quantum mechanics informs the atomic structure of matter, a universe does not grow smoothly by rescaling $a(\tau)$ as time goes on. Instead, space can only be enlarged by incremental excitations of geometry, much like one would generate a macroscopic field by creating a large number of single particles. In this way our current large universe has to emerge from a small quantum state at the big bang. The growth of such a discrete universe is not easy to describe, but several general properties of such scenarios can be found, e.g., in [7–11]. While such discrete steps are not noticeable today, they would have been crucial at small sizes of a very early universe, especially at the big bang singularity. Moreover, here it becomes important that a discrete space has only a finite amount of storage for energy due to a lower limit for wavelengths of supported waves. An upper limit for energy densities results, in contrast to the lack of an upper bound in classical physics. In quantum gravity, some deviation from classical physics has to happen at high energies as we approach the big bang.

8.3.1 Loop Quantum Cosmology

Space-time is not only atomic but also quantum when quantum gravity is invoked. For isotropic cosmology, the scale factor thus becomes on operator acting on wave functions of a universe. If we use the square $p = a^2$ of the scale factor as the basic variable (as suggested by isotropic loop quantum cosmology [1, 12]), its canonical momentum is proportional to simply the proper time derivative $P = \dot{a}$. In a Schrödinger representation, this would give us operators $\hat{p} = \hat{a}^2$ acting by multiplication and $\hat{P} = -i\hbar\partial/\partial a^2$. (We write a partial derivative since there may be other variables such as a scalar field ϕ.)

This quantum representation would imply that a is continuous since its operator has a continuous spectrum and its momentum operator is an infinitesimal translation. In such a (Wheeler–DeWitt [13, 14]) quantization there would be no discrete spatial growth, which rather can only come from finite shift operators such as a quantization of $\exp(if(p)P)$. Here, $f(p)$ takes into account a possible dependence of the discreteness scale (and thus the allowed shift sizes) on the total size. Our basic variable p would be equidistantly spaced for $f(p) = \text{const}$, but other possibilities are certainly possible. A precise form would have to be derived from a concrete model of the refinement processes going on in a quantum universe. At the required fundamental level, this is currently not manageable, but general arguments provide restrictions. If $f(p) \propto p^x$ is assumed to be a power law, for instance, then loop quantum gravity as presently understood can only accommodate powers in the range $-1/2 < x < 0$ [15], with several independent phenomenological and stability arguments preferring a value near the lower bound $x \approx -1/2$ [5, 16–20].

Cosmological models in loop quantum cosmology [1] implement the discreteness and thus represent only finite shift operators in p. To express the dynamical law in terms of loop variables, one thus has to replace $\dot{a}^2$ in the Friedmann equation (8.1) by a term involving only $\exp(if(p)P)$ in addition to a itself. There cannot be an exact correspondence since there is no operator for P (or a logarithm of the quantized exponential), such that higher order terms in extrinsic curvature P result from a series expansion. The form of such higher order terms is dictated by the quantization, but is not much restricted in homogeneous models due to quantization ambiguities; stronger consistency conditions result when homogeneous models are embedded in inhomogeneous ones. For instance, we may quantize the Friedmann equation (or, more precisely, the Hamiltonian constraint which follows after multiplying with a^3) as

$$\left(f(p)^{-2}\widehat{\sin^2(f(p)P)\sqrt{p}}\right)\psi(p,\phi) = -\frac{4\pi G\hbar^2}{3}\widehat{a^{-3}}\frac{\partial^2}{\partial\phi^2}\psi(p,\phi)\,. \qquad (8.4)$$

This equation has by construction the correct classical limit. On the left hand side, we are using a shift operator such that it presents a difference equation [21]. The factor ordering between $f(p)$ and P is to be decided, which we will do below. Moreover, on the right hand side we have to define the operator quantizing a^{-3}. Also here, there are traces of the spatial discreteness because $\hat{a}$ in loop quantum cosmology has a discrete spectrum containing zero, and thus lacks a densely defined inverse. Nevertheless, one can find operators which have a^{-3} as the classical limit and which are finite [22] even at zero eigenstates of $\hat{a}$. These constructions make use of techniques first developed in the full theory of loop quantum gravity [23] and thus faithfully reflect the spatial discreteness.

8.3.2 *Harmonic Cosmology*

Our main interest here is in states which become semiclassical at large volume. As one evolves backwards to smaller volume, stronger quantum properties can easily set in, but initially the factor ordering in the quantum constraint should not play a large role. Whether it does so later on, once small volume is reached, can be checked self-consistently from solutions. For large volume, moreover, deviations of expectation values $\langle\widehat{a^{-3}}\rangle$ from the classical behavior of a^{-3} can be ignored. There is then one final problem before one can analyze a difference equation of quantum cosmology: there is no absolute time which one could use to follow Hamiltonian evolution. This problem, however, is trivially solved for a free, massless scalar where the scalar ϕ itself is monotonic and can play the role of time. Multiplying with a^3, assuming $\hat{a}^3\widehat{a^{-3}} \approx 1$ and taking a square root in (8.4), we easily obtain an evolution equation in the form of a Schrödinger equation:

$$\hat{p}_\phi\psi(p,\phi) = i\hbar\frac{\partial}{\partial\phi}\psi(p,\phi) = \pm\sqrt{\frac{3}{4\pi G}}\left|f(p)^{-1}\widehat{\sin(f(p)P)p}\right|\psi(p,\phi)\,. \quad (8.5)$$

In general, the sign on the right hand side in combination with the absolute value of an operator is a subtle issue. The absolute value means that only positive energy solutions are considered which, due to the freedom of sign, can appear in superpositions of left- and right-moving states. Ignoring the absolute value, in general, would mean that the wrong types of states would be allowed in superpositions, with potentially fatal consequences for the correct evolution.

But this turns out to be unproblematic for the states we are interested in: Having a state which is semiclassical at least at one instant of large volume requires matter fluctuations to be small compared to the total value: $\Delta p_\phi \ll |p_\phi|$. Since p_ϕ is a constant of motion, this condition is not only satisfied once but throughout the evolution, even if the state may become highly non-semiclassical (i.e., other fluctuations could become large). We are thus dealing with states only which are sharply peaked at a large, macroscopic value of $|p_\phi|$. The overwhelming part of such states, as superpositions of $\hat{p}_\phi$-eigenstates, is supported at only one sign of p_ϕ, and any spill-over to the opposite sign can be neglected. In fact, one can project out the negative frequency contributions of an initial state, which then remains supported only on positive frequency eigenstates.[1] In this way, the absolute value becomes an issue of choosing appropriate initial states, rather than an issue of evolution. Of main interest in our analysis will be expectation values and fluctuations of volume and curvature, and we will only have to make sure that initial values of those variables we choose can consistently be realized by a positive frequency state. One can easily convince oneself that this is possible, for instance by projecting off the negative frequency contributions from a Gaussian state. For $\Delta p_\phi \ll |p_\phi|$, the projection does not significantly change expectation values and fluctuations of operators which are insensitive to $p_\phi = 0$.

One may view this issue as a finite size effect: The wave function in energy space would have to be projected to the positive energy axis, removing part of its tail at negative energies. This is comparable to treating a particle in a cavity (e.g., an elementary particle in a detector) as a wave packet, which would have to be projected to the interior region of the cavity. As is well known, such effects can be ignored completely unless the particle is close to the boundary. What is more important, and in fact crucial, is possible interactions of the particle with other fields, which can be included by suitable terms in an effective theory. Similarly, the condition $\Delta p_\phi \ll |p_\phi|$ ensures that our system, in energy space, is far away from its boundary and the boundary effect – positivity of p_ϕ – can safely be ignored. In what follows we will determine whether there are complicated interactions whose inclusion in solutions would be much more crucial.

It turns out that one can exploit the freedom in the factor ordering to bring the system in explicitly solvable form, corresponding to a free theory [6]. This fact has allowed many new, sometimes surprising derivations of properties of solutions as well as dynamical coherent states for this system. To make this explicit, we first perform a canonical transformation from our pair (p, P) to a new pair $(V, f(p)P)$ with $V := p/f(p)$, for $f(p) \propto p^x$ with some x. Notice that V is the spatial volume for the power-law case $f(p) \propto p^{-1/2}$, but it can take other values depending on x. The fact that x near $-1/2$ seems preferred implies that for dynamical purposes of an isotropic universe the volume should be considered as nearly equidistantly spaced. As in this example, phenomenological considerations can provide detailed insights into the fundamental form of quantum space-time.

[1] This projection may be difficult to perform explicitly, especially since it would require non-local representations of the Hamiltonian operator in a p or P-representation of states. The treatment used here does not require one to deal with any such complications.

To realize the solvability of the model [6], the new canonical pair still has to be transformed to non-canonical basic variables V in combination with $J := V\exp(if(p)P)$. Quantum analogs of these basic variables, with $\hat{J}$ ordered as in the definition, satisfy the commutation relations

$$[\hat{V},\hat{J}] = \hbar\hat{J}\,,\quad [\hat{V},\hat{J}^\dagger] = -\hbar\hat{J}^\dagger\,,\quad [\hat{J},\hat{J}^\dagger] = -2\hbar\hat{V} - \hbar^2 \tag{8.6}$$

of an sl$(2,\mathbb{R})$ algebra. Moreover, we can express our Hamiltonian as a linear combination of basic operators, $\hat{H} = -\frac{1}{2}i(\hat{J} - \hat{J}^\dagger)$ (ignoring factors of $4\pi G/3$ and $\hbar$), which specifies a suitable factor ordering. By (8.5), the action of $\hat{H}$ on a wave function ψ has to equal $\hat{p}_\phi\psi$. This is our dynamical equation to solve for properties of states.

With a linear Hamiltonian in a linear algebra of basic operators, our system is easily solvable. Instead of taking a detour of computing complete states, which would carry much more information than needed for our purposes, we can directly derive equations of motion for expectation values and find their solutions. Thanks to the linearity, the equations of motion

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\phi}\langle\hat{V}\rangle &= \frac{\langle[\hat{V},\hat{H}]\rangle}{i\hbar} = -\frac{1}{2}(\langle\hat{J}\rangle + \langle\hat{J}^\dagger\rangle),\\ \frac{\mathrm{d}}{\mathrm{d}\phi}\langle\hat{J}\rangle &= \frac{\langle[\hat{J},\hat{H}]\rangle}{i\hbar} = -\langle\hat{V}\rangle - \frac{1}{2}\hbar = \frac{\mathrm{d}}{\mathrm{d}\phi}\langle\hat{J}^\dagger\rangle\end{aligned} \tag{8.7}$$

are not coupled to fluctuations or higher moments and can be solved directly without knowing further properties of the spreading of states. This gives

$$\langle\hat{V}\rangle(\phi) = \frac{1}{2}(Ae^{-\phi} + Be^{\phi}) - \frac{1}{2}\hbar \tag{8.8}$$

$$\langle\hat{J}\rangle(\phi) = \frac{1}{2}(Ae^{-\phi} - Be^{\phi}) + iH \tag{8.9}$$

with two constants of integration A and B and $H = \langle\hat{H}\rangle$.

Using partially complex variables, we also have to impose reality conditions to make sure that operators we implicitly use in our equations of motion have the correct adjointness relations, and that states are physically normalized. We do not use explicit states, but expectation values of basic operators in physically normalized states must show the correct adjointness behavior: $\hat{V}$ is self-adjoint, while $\hat{J}$ has to satisfy $\hat{J}\hat{J}^\dagger = \hat{V}^2$ to ensure unitarity of the quantized exponential. Taking expectation values implies that $\langle\hat{V}\rangle$ must be real, which is easy to impose, and, due to $\hat{J}\hat{J}^\dagger = \hat{V}^2$, both expectation values must be related to the fluctuation $(\Delta V)^2 = \langle\hat{V}^2\rangle - \langle\hat{V}\rangle^2$ and the covariance $C_{J\bar{J}} = \frac{1}{2}\langle\hat{J}\hat{J}^\dagger + \hat{J}^\dagger\hat{J}\rangle - \langle\hat{J}\rangle\langle\hat{J}^\dagger\rangle$ by

$$|\langle\hat{J}\rangle|^2 - \left(\langle\hat{V}\rangle + \tfrac{1}{2}\hbar\right)^2 = (\Delta V)^2 - C_{J\bar{J}} + \frac{1}{4}\hbar^2\,. \tag{8.10}$$

For a state which is semiclassical at least once, i.e. ΔV and $C_{J\bar{J}}$ are of the order $\hbar$ for one value of ϕ, we obtain $AB = H^2 + O(\hbar)$ by inserting our solutions in (8.10). Under this condition, A and B must have the same sign, such that we can define $A/B =: e^{2\delta}$ and the solution $\langle\hat{V}\rangle(\phi) = H\cosh(\phi - \delta)$ (ignoring the small contribution $-\frac{1}{2}\hbar$) demonstrates that no singularity of vanishing volume is reached. Instead, the state bounces off a minimal volume at $\phi = \delta$ and enters a new, classically invisible region of space-time.

For $x = -1/2$, i.e., when V is indeed the spatial volume, it is straightforward to translate the internal time solution $\langle\hat{V}\rangle(\phi)$ into a proper-time expression. In this case, the scalar momentum is given by[2] $H = p_\phi = a^3\dot{\phi} = V\dot{\phi} = H\cosh(\phi-\delta)\dot{\phi}$ in terms of the proper time derivative of ϕ. The scalar is thus simply related to proper time τ by $\phi(\tau) = \delta + \mathrm{arsinh}\tau$ (for $\tau = 0$ at the bounce), and we have

$$\langle\hat{V}\rangle(\tau) = H\sqrt{1+\tau^2}\,. \tag{8.11}$$

This clearly shows the transition between the classical behavior $a^3 \propto \tau$ as in the solution (8.3) and non-classical behavior at the bounce for $\tau = 0$, as illustrated in Fig. 8.1.

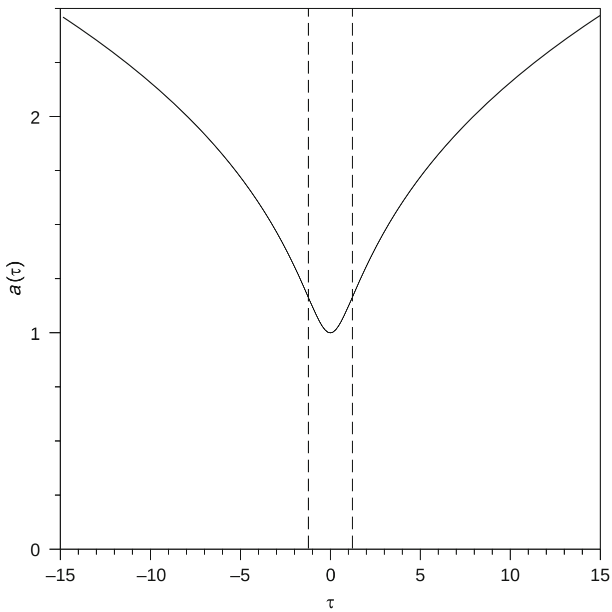


Fig. 8.1 Scale factor (in units of $p_\phi^{1/3}$) as a function of proper time for an isotropic free massless scalar model in loop quantum cosmology. The brief phase in which gravity is repulsive is indicated by the dashed lines

[2] The proper time derivative $\dot{\phi}$ follows from the Hamiltonian constraint by $\dot{\phi} = \{\phi, H_{\mathrm{matter}}\}$. In our case, we do not change the matter term $H_{\mathrm{matter}} = \frac{1}{2}a^{-3}p_\phi^2$ by quantum corrections, such that we can use the classical relation to relate $\dot{\phi}$ to p_ϕ, but with the effective solution $a^3 = \langle\hat{V}\rangle$.

8.3.2.1 Effective Friedmann Equation

For a linear system, one can easily derive precise effective equations whose solutions determine the evolution of expectation values. Such equations are usually more direct and intuitive to interpret, compared to equations for wave functions. In our solvable model for a quantum space-time, we have the effective Hamiltonian $H_{\rm eff} = \langle \hat{H} \rangle = (2i)^{-1}(J - \bar{J}) = V \sin \bar{P} = p_\phi$ (using $\bar{P} := f(p)P$, dropping brackets for expectation values as in $V = \langle \hat{V} \rangle$ and ignoring factor ordering terms), and the equation of motion

$$\frac{\mathrm{d}V}{\mathrm{d}\phi} = -\frac{J + \bar{J}}{2} = -V \cos \bar{P} = -V \sqrt{1 - \left(\frac{p_\phi}{V}\right)^2}$$

where we have eliminated $\bar{P}$ via the effective Hamiltonian which equals p_ϕ. (At this stage we also make use of reality conditions since $\bar{P}$ must be real for trigonometric identities to apply.)

With $\dot{\phi} = V^{-3/2(1-x)} p_\phi$, this implies

$$\dot{V}^2 = \left(\frac{\mathrm{d}V}{\mathrm{d}\phi}\right)^2 \dot{\phi}^2 = p_\phi^2 V^{-(1+2x)/(1-x)} \left(1 - \left(\frac{p_\phi}{V}\right)^2\right)$$

for the derivative by proper time, and thus the corrected Friedmann equation

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{1}{2(1-x)^2} \rho_\phi (1 - 2a^{2+4x} \rho_\phi)$$

for a free, massless scalar. In this form, the quantum correction can simply be formulated as a term quadratic in the energy density of the free scalar [24]. One should note, however, that this is not the primary correction in loop quantum cosmology. The reason for the new term is the higher curvature corrections in the effective Hamiltonian, which in this model can be reformulated as a simple correction of the energy dependence. This proves the correctness of this effective equation[3] [6], but only for this specific matter content. We will later discuss which additional correction terms one has to expect in general, such as for a massive or self-interacting scalar.

[3] The right hand side of the effective Friedmann equation is a-independent only for $x = -1/2$, and the classical scale-invariance of solutions is preserved only in this case. Still, other values for x are possible since there is no need to preserve the accidental classical rescaling symmetry of spatially flat Friedmann–Robertson–Walker models. This does not present a (diffeomorphism) gauge problem since diffeomorphisms are trivial in the reduced homogeneous model. The meaning of the scale dependence as a trace of the underlying discreteness scale becomes clear when embedding the homogeneous models in inhomogeneous ones [15].

8.3.2.2 Cosmic Forgetfulness

The procedure can be repeated for higher moments of a state. Just like expectation values in (8.7), fluctuations and higher moments of a state are dynamical. For a solvable system, we have a finite number of coupled equations at any order, e.g., for fluctuations and correlations. Fluctuations and correlations satisfy equations of motion

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\phi}(\Delta V)^2 &= -C_{VJ} - C_{V\bar{J}}, \frac{\mathrm{d}}{\mathrm{d}\phi}(\Delta J)^2 = -2C_{VJ}, \frac{\mathrm{d}}{\mathrm{d}\phi}(\Delta \bar{J})^2 = -2C_{V\bar{J}} \\
\dot{C}_{VJ} &= -\frac{1}{2}(\Delta J)^2 - \frac{1}{2}C_{J\bar{J}} - (\Delta V)^2, \dot{C}_{V\bar{J}} = -\frac{1}{2}(\Delta \bar{J})^2 - \frac{1}{2}C_{J\bar{J}} - (\Delta V)^2 \\
\dot{C}_{J\bar{J}} &= -C_{VJ} - C_{V\bar{J}}
\end{aligned} \tag{8.12}$$

and are subject to uncertainty relations [25]. Solutions for fluctuations,

$$(\Delta V)^2(\phi) = \frac{1}{2}(c_3e^{-2\phi} + c_4e^{2\phi}) - \frac{1}{4}(c_1 + c_2) \tag{8.13}$$

$$(\Delta J)^2(\phi) = \frac{1}{2}(c_3e^{-2\phi} + c_4e^{2\phi}) + \frac{1}{4}(3c_2 - c_1) - i(c_5e^{\phi} - c_6e^{-\phi}) \tag{8.14}$$

$$(\Delta \bar{J})^2(\phi) = \frac{1}{2}(c_3e^{-2\phi} + c_4e^{2\phi}) + \frac{1}{4}(3c_2 - c_1) + i(c_5e^{\phi} - c_6e^{-\phi}) \tag{8.15}$$

$$C_{VJ}(\phi) = \frac{1}{2}(c_3e^{-2\phi} - c_4e^{2\phi}) + \frac{i}{2}(c_5e^{\phi} + c_6e^{-\phi}) \tag{8.16}$$

$$C_{V\bar{J}}(\phi) = \frac{1}{2}(c_3e^{-2\phi} - c_4e^{2\phi}) - \frac{i}{2}(c_5e^{\phi} + c_6e^{-\phi}) \tag{8.17}$$

$$C_{J\bar{J}}(\phi) = \frac{1}{2}(c_3e^{-2\phi} + c_4e^{2\phi}) + \frac{1}{4}(3c_1 - c_2) \tag{8.18}$$

can be chosen to saturate uncertainty relations, thus determining properties of dynamical coherent states. Of particular interest are volume fluctuations $(\Delta V)^2(\phi) \approx \hbar H \cosh(2(\phi - \delta_2))$ in the saturating case where in general $\delta_2 \neq \delta$ appearing in $\langle \hat{V} \rangle(\phi)$. Thus, fluctuations need not be symmetric around the bounce point. In terms of proper time τ we have, for $x = -1/2$,

$$(\Delta V)^2(\tau) = \hbar H \left(\cosh(2\delta_2)(1 + 2\tau^2) - 2\sinh(2\delta_2)\tau\sqrt{1 + \tau^2}\right) \tag{8.19}$$

which is illustrated in Fig. 8.2. In this way, solutions easily demonstrate the control, or lack thereof, one has on fluctuations before the big bang.

In fact, what this procedure easily shows is that there was a space-time before the big bang: the classical boundary presented by the big bang singularity is eliminated by switching on quantum-mediated repulsive contributions to the gravitational force. In general, however, this new patch of the world is truly a quantum space-time whose quantum variables such as fluctuations and correlations present additional

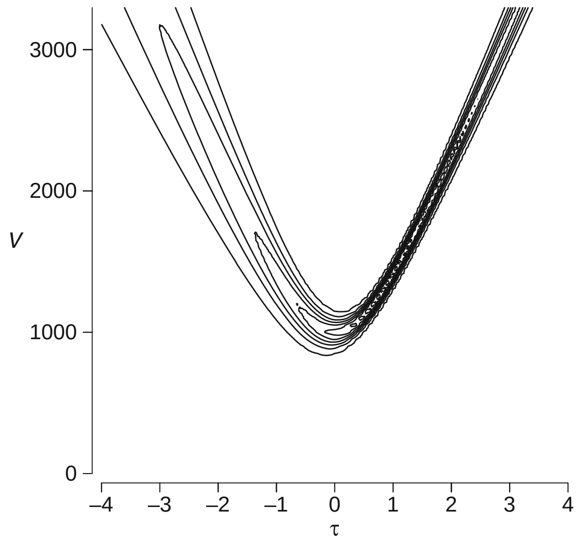


Fig. 8.2 Contour plot of a generic wave function with spread (8.19) and expectation value (8.11), whose fluctuations before and after the bounce can deviate from each other by factors different from one

degrees of freedom to be reckoned with. These variables become important, in addition to the expectation values of a state which one can identify with components in a classical line element. They can become large when the state travels through a deep quantum regime such as the big bang. Even in models where the volume does not shrink down all the way to Planck size, which could make one expect that quantum effects are not too strong, bounds on quantum fluctuations before the big bang based on the fact that we have a very nearly classical state at large volume are rather weak [3].

For dynamical coherent states, one can use the saturation equations to solve for all parameters c_I of (8.13–8.18) in terms of the constant A (assumed to equal B by shifting the origin of ϕ) in the solution for $\langle\hat{V}\rangle(\phi)$, which determines the size of the universe at the bounce, and the matter Hamiltonian H and its fluctuation ΔH. In particular, one can solve for the asymmetry

$$D := \left|\lim_{\phi\to-\infty}\frac{(\Delta V)^2}{\langle\hat{V}\rangle^2} - \lim_{\phi\to\infty}\frac{(\Delta V)^2}{\langle\hat{V}\rangle^2}\right| = 2\frac{|c_3-c_4|}{A^2} \tag{8.20}$$

$$= 4\frac{H}{A}\sqrt{\left(1-\frac{H^2}{A^2}+\frac{1}{4}\frac{\hbar^2}{A^2}\right)\frac{(\Delta H)^2}{A^2}-\frac{1}{4}\frac{\hbar^2}{A^2}+\left(\frac{H^2}{A^2}-1\right)\frac{(\Delta H)^4}{A^4}} \tag{8.21}$$

of volume fluctuations around the bounce [4]. Of more interest is this quantity divided by c_4 because it is insensitive to the total size of fluctuations:

$$\left|1-\frac{(\Delta V)^2_-}{(\Delta V)^2_+}\right| = \frac{|c_4-c_3|}{c_4}$$

$$= \frac{2\frac{H}{A}\delta}{\frac{\delta^2}{2(\Delta H)^2} \pm \frac{H}{A}\delta + \frac{1}{2}\frac{H^2}{A^2}(\Delta H)^2 + \frac{1}{8}\frac{A^2\hbar^2}{(\Delta H)^2}} \qquad (8.22)$$

where

$$\delta := \sqrt{\left(\frac{H^2}{A^2} - 1\right)(\Delta H)^4 + \left(A^2 - H^2 + \frac{1}{4}\hbar^2\right)(\Delta H)^2 - \frac{1}{4}A^2\hbar^2}\,.$$

Estimating the right hand side for typical semiclassical states shows that it need not be small and can easily be of the order ten or so. Thus, generic states can be very asymmetric around the bounce.[4]

Interestingly, it is exactly the large bounce volume (or large H) in this case which makes states before the big bang highly sensitive to initial conditions one could pose after the big bang for a backward calculation [4], making the universe forget some of its pre-bounce past. For smaller H, on the other hand, one would certainly enter a much stronger quantum regime. Thus, in either case – with large bounce sizes or small ones – no precise knowledge about the state of the universe before the big bang results. There is no avoiding the fact that quantum regimes do naturally lead to strong quantum states, and thus quantum space-time.

8.3.3 Interactions: How Quantum Is The Bounce?

When the model deviates from an isotropic free scalar model, be it by a matter potential, anisotropies or inhomogeneities, higher moments of the state couple to expectation values. We are no longer dealing with a free theory in such a

[4] It has been claimed in [26] that states have to be very nearly symmetric, based on a much weaker estimate for the asymmetry. In fact, it was shown that the difference of relative fluctuations had to be small as an absolute number, but this does not mean much unless one knows the size of each term in the difference. It is easy to show that the numerical example provided in [26] easily allows factors between volume fluctuations before and after the bounce as large as 10^{28} (!) which can hardly be called symmetric. The estimate does not even show whether fluctuations before the bounce can be considered semiclassical to the same degree to which they are semiclassical after the bounce. The reason for the extreme weakness of the inequality of [26] is the fact that the difference of *quadratic* relative fluctuations is bounded by a *linear* fluctuation term. For small numbers such as relative fluctuations, squares are certainly much smaller than the original numbers. This makes each term in the difference defining the asymmetry very small, not just the difference itself. The relation (8.22) used here is much stronger because it is not sensitive to the absolute size of fluctuations. Although it applies to the more special class of dynamical coherent states, which retain their degree of semiclassicality at all times in the sense of saturating uncertainty relations, significant asymmetries in volume fluctuations are still possible.

case, and state properties do influence the effective motion. For small evolution intervals starting from a semiclassical state, such quantum corrections remain small. However, cosmology deals with long evolution times, such that quantum states can change dramatically and lead to stronger quantum corrections. In particular for a decision of whether the universe still bounces in this case one has to evolve from a semiclassical state at large volume all the way to where the bounce might occur. While a semiclassical state starting near the bounce would safely travel through the bounce nearly as smoothly as in the free model, a state coming from large volume may have changed significantly. For such a state, equations of motion would have strong quantum corrections and a reliable statement about the bounce – whether it persists or, if it does, what its precise properties are – is much more complicated than in the free case.

The case of a scalar potential has already been studied [27], although no complete analysis is available especially regarding the bounce. With a potential, it is no longer possible to use ϕ as an internal time globally. But one can analyze the local behavior and see what implications the interaction effects may have. The Friedmann equation now implies a Hamiltonian

$$p_\phi = H = |p|\sqrt{c^2 - |p|W(\phi)} \tag{8.23}$$

for evolution in ϕ, which is no longer quadratic or linear. (To be specific, we now use $x = 0$, again calling $a^2 = p$ instead of V. Moreover, we denote the scalar potential as $W(\phi)$ to distinguish it from the volume.) In addition to the internal time difficulty, the ϕ-dependence of the potential implies that p_ϕ and H no longer commute after quantization, such that the first order Schrödinger-type equation does not produce exact solutions of the original second order equation for p_ϕ^2 of Klein–Gordon type. Nevertheless, for a small potential, which then also allows perturbative treatments, one can see that despite those two issues much information about the behavior can be gained [27].

The effective equations in this case contain explicit coupling terms between expectation values and fluctuations (or higher moments which, however, can be ignored compared to fluctuations in an initial approximation of semiclassical states). For $\langle\hat{p}\rangle$, we obtain

$$\begin{aligned}
\frac{\mathrm{d}\langle\hat{p}\rangle}{\mathrm{d}\phi} &= -\frac{\langle\hat{J}\rangle + \langle\hat{J}^\dagger\rangle}{2} + \frac{\langle\hat{J}\rangle + \langle\hat{J}^\dagger\rangle}{(\langle\hat{J}\rangle - \langle\hat{J}^\dagger\rangle)^2}\langle\hat{p}\rangle^3 W(\phi) \\
&\quad +3\frac{\langle\hat{J}\rangle + \langle\hat{J}^\dagger\rangle}{(\langle\hat{J}\rangle - \langle\hat{J}^\dagger\rangle)^4}\langle\hat{p}\rangle^3((\Delta J)^2 + (\Delta\bar{J})^2 - 2C_{J\bar{J}})W(\phi) \\
&\quad -6\frac{\langle\hat{J}\rangle + \langle\hat{J}^\dagger\rangle}{(\langle\hat{J}\rangle - \langle\hat{J}^\dagger\rangle)^3}\langle\hat{p}\rangle^2(C_{pJ} - C_{p\bar{J}})W(\phi) \\
&\quad +3\frac{\langle\hat{J}\rangle + \langle\hat{J}^\dagger\rangle}{(\langle\hat{J}\rangle - \langle\hat{J}^\dagger\rangle)^2}\langle\hat{p}\rangle(\Delta p)^2\, W(\phi)
\end{aligned}$$

$$
\begin{aligned}
&-\frac{2\langle\hat{p}\rangle^3}{(\langle\hat{J}\rangle-\langle\hat{J}^\dagger\rangle)^3}((\Delta J)^2-(\Delta\bar{J})^2)W(\phi)\\
&+\frac{3\langle\hat{p}\rangle^2}{(\langle\hat{J}\rangle-\langle\hat{J}^\dagger\rangle)^2}(C_{pJ}+C_{p\bar{J}})W(\phi)\,.
\end{aligned}
\tag{8.24}
$$

As one can easily see, state properties are now required to find solutions for expectation values. If one chooses an initial semiclassical state, all the fluctuations and correlations are small compared to the classical correction implied by the potential. Thus, the expectation value will initially follow the free trajectory closely with the only correction coming from the potential, assumed to be small in the derivation of (8.24). In particular, if one starts near the free bounce with a semiclassical state, the state will, as anticipated, bounce in a very similar way even in the presence of quantum corrections. However, one cannot reliably assume that the state near the bounce is semiclassical. One has to know how fast the state, safely assumed to be semiclassical *at large volume*, will start to deviate from being semiclassical, implying that fluctuation terms will have to be taken into account. Especially for the bounce, what is important is not whether a state which is semiclassical near the bounce will bounce, but rather whether a state which starts out semiclassically at large volume – very far from the bounce – will still bounce and in which way, even though generically it would no longer be semiclassical in this regime.

Such questions, which are typical for the long evolution times involved in cosmology, are difficult to address because assumptions made so far prevent one from doing a long-term analysis. But there is potential for a numerical analysis of the effective equations, where one can, for instance, avoid the internal time problem by patching different solutions. Only such an investigation can tell us whether the classical singularity is replaced by a smooth bounce, as the free model suggests, or generically involves a strong quantum phase.

There is some preliminary information available on how state properties can matter. In the free model, we know properties of dynamical coherent states as derived in [25]. They can be used as zeroth order solutions, to be corrected by perturbative terms in the presence of a potential. While the effective equations such as (8.24) in general involve several terms of similar orders, near the bounce one can use several simplifying assumptions: we have $\langle\hat{J}\rangle+\langle\hat{J}^\dagger\rangle\approx 0$ and $\langle\hat{p}\rangle\approx\langle\hat{H}\rangle=-\frac{1}{2}i(\langle\hat{J}\rangle-\langle\hat{J}^\dagger\rangle)$. In the solvable model this implies $\mathrm{d}\langle\hat{p}\rangle/\mathrm{d}\phi\approx 0$, as it should clearly happen at the bounce. With a perturbative scalar potential, most terms in the perturbation equation for $\mathrm{d}\langle\hat{p}\rangle/\mathrm{d}\phi$ drop out as they contain $\langle\hat{J}\rangle+\langle\hat{J}^\dagger\rangle$. The last two terms of (8.24), however, remain and give

$$
\frac{\mathrm{d}\langle\hat{p}\rangle}{\mathrm{d}\phi}\approx\left(\frac{1}{2}\mathrm{Im}(\Delta J(\phi))^2-\frac{3}{2}\mathrm{Re}C_{pJ}(\phi)\right)W(\phi)\,.
$$

Based on the dynamical coherent state properties of [25], this is zero at $\phi\sim 0$ for special states which are unsqueezed (i.e., $c_3\sim c_4$, $c_5\sim c_6$), but non-zero for squeezed states. If no strong squeezing develops during the evolution from large volume to near the bounce, not much should happen to the free bounce to the order of approximations used here. If the state develops strong squeezing or is squeezed already as an initial semiclassical state, however, the bounce behavior must change: at the free bounce, i.e., where $\mathrm{d}\langle\hat{p}\rangle/\mathrm{d}\phi = 0$ in the free model, $\mathrm{d}\langle\hat{p}\rangle/\mathrm{d}\phi$ will be non-zero in the interacting system. The universe may certainly bounce elsewhere, but to check this further analysis is required. Moreover, as per cosmic forgetfulness, no strong constraint on c_3/c_4 or c_5/c_6 can be assumed: For all we know, the state of the universe may be strongly squeezed, making the bounce sensitive to quantum corrections.

8.3.4 Black Holes

Next to cosmological singularities, black holes provide further examples where the classical space-time reaches a boundary in finite time. For a black hole space-time, inhomogeneities are strictly required, but there is one case, the interior inside the horizon of a non-rotating black hole in vacuum, where one can use a homogeneous description. In this region, the Schwarzschild solution

$$\mathrm{d}s^2 = -\left(1 - \frac{2M}{r}\right)\mathrm{d}t^2 + \left(1 - \frac{2M}{r}\right)^{-1}\mathrm{d}r^2 + r^2\mathrm{d}\Omega^2$$

turns into a homogeneous model

$$\mathrm{d}s^2 = -\left(\frac{2M}{T} - 1\right)^{-1}\mathrm{d}T^2 + \left(\frac{2M}{T} - 1\right)\mathrm{d}R^2 + T^2\mathrm{d}\Omega^2$$

(with $T = r$ and $R = t$) of Kantowski–Sachs type because the static symmetry, which is timelike outside the horizon ($r > 2M$), turns spacelike inside the horizon ($r < 2M$) and provides the additional spatial symmetry extending the rotational one to a transitive group action: Coefficients in the metric now are time rather than space dependent.

Since the interior is anisotropic, the gravitational equations are not free and receive quantum corrections from coupling terms between expectation values and fluctuations as in our isotropic model of Section 8.3.3. In contrast to cosmological models, here we do not have a large parameter at our disposal, by which we could push all geometrical variables to large values. In homogeneous cosmology, this large parameter, p_ϕ, is the matter content of the whole universe; but the black hole interior is homogeneous only in the vacuum case. Thus, we do not have the luxury of large sizes and have to deal with tiny, squashed geometries near the classical singularity.

In particular, if bounces were to happen, they could only occur at small extensions, which for astrophysical black holes of large masses would be reached in long evolution times after crossing the horizon. The situation here is more complicated than in interacting isotropic models of Section 8.3.3 since quantum effects not only have time to grow, but do not even need to become very large to be comparable to classical variables near the singularity. Thus, in this case the situation of possible bounces is much less certain. ("Effective" equations which include some higher curvature corrections but ignore quantum back-reaction have been studied in several articles [28–35] for the Schwarzschild interior as well as anisotropic cosmological models. Not surprisingly, these solutions do exhibit bounces. Because quantum back-reaction was completely ignored, however, the significance of these results remains unclear.)

Instead of using effective equations, one can discuss the structure of such black holes in quantum gravity at a more fundamental level, directly for wave functions. This does not easily provide intuitive pictures of the space-time geometry and its possible extension beyond the classical singularity, but can nevertheless result in the formulation of general expectations. Since the near-singularity region must correspond to a strong quantum regime, one should study the fundamental equation for a quantum state directly. One will have to face complicated issues of interpreting the wave function in quantum gravity, but generic statements can nevertheless be made. In particular, if one can show that the wave function as the fundamental object in this framework extends beyond the classical singularity, space-time is extended no matter how the wave function is to be interpreted; see also [36] for a detailed discussion of this issue of quantum hyperbolicity.

For the black hole interior in loop quantum gravity, the unique extension of the wave function beyond the classical singularity was demonstrated in [37, 38], following the example of isotropic and homogeneous cosmology [39–41]. The interior must then extend to a region beyond the classical singularity, which has to be matched to some exterior space-time. To find the precise matching, inhomogeneous space-times in spherical symmetry have to be understood, which is certainly more difficult than the homogeneous treatments [42, 43] (see also [44, 45] for the related case of Gowdy models). Based on available indications, however, a picture has been suggested [46] where the new region matches to the old exterior, rather than splitting off into a new, disconnected region of space-time. This is also consistent with an analysis of matched space-time solutions obtained from effective equations of collapsing matter [47]. A detailed discussion of those issues can be found in [48].

To decide whether the interior reconnects to the old exterior after the singularity has been crossed, one certainly needs a handle on the inhomogeneous situation. The basic statement of extendability of the wave function has indeed been shown to be realized in spherical symmetry – without the homogeneity assumption [49]. Also other properties such as a well-defined behavior of some curvature variables can be derived in the spherically symmetric case [50]. Moreover, several proposals to study the evolution of space-time near a spherically symmetric horizon have been made and are being developed further [51–54]. However, understanding the endpoint of evaporation in this way, which would be necessary to decide whether the interior reconnects at this point, requires much more work.

In principle, one can imagine that the interior separates from the exterior even in canonical quantum evolution, where topology change is often deemed impossible. The space-time atoms simply are being ripped apart at one point, which can in fact be realized in a precise way in loop quantum gravity: the metric (or triad) would degenerate at one point, allowing adjacent space-time regions to separate off. What would make this behavior difficult to interpret as a splitting off of the interior is the fact that such a separation of space-time atoms, if it can occur at one point in the strong curvature region, could easily appear everywhere along the classical singularity. In other words, no known mechanism exists which would prevent quantum space-time to fragment into disconnected space-time atoms if it is supposed to split into two disconnected parts at one point. If the entire space-time remains intact and the interior reconnects to the old exterior, on the other hand, there is no such potential inconsistency. Interestingly, the resulting picture of a non-singular evaporating black hole, as illustrated in Fig. 8.3, agrees with other proposals based on properties of quantum energy-momentum in general relativity [55, 56].

Another supporting argument for this picture is the symmetry of wave functions describing the interior geometry under time reflection along the classical singularity. This can be seen as a consequence of consistency conditions imposed by the difference equation which describes the dynamics [57].[5] Thus, the new quantum space-time region after crossing the black hole singularity is simply a time reversal of the classical interior. The fact that there are no crucial changes in the geometry, even after traversing the strong curvature region, shows that one can consistently match the extended interior to a static exterior as it must be realized according to the Birkhoff theorem. The fact that properties of difference equations describing a homogeneous

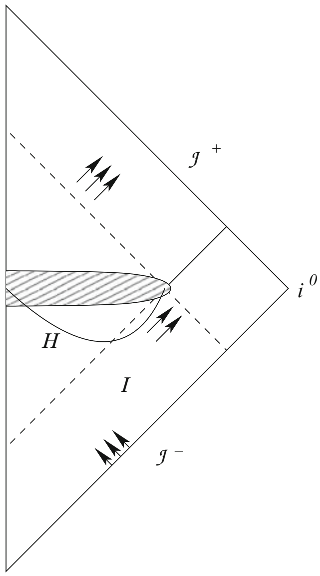


Fig. 8.3 Penrose diagram of a non-singular black hole space-time. The classical singularity merely presents a high curvature interior region. Not only Hawking radiation escapes along the horizon, but also infalling matter after traversing the classical singularity [46]

[5] Technically, this is analogous to relations which can be used for dynamical initial conditions of a wave function of a universe [58, 59].

quantum interior are consistent with a classical result about spherically symmetric space-times is rather non-trivial and encouraging. More detailed studies of the underlying difference equations are under way; see e.g., [17, 60, 61].

8.4 Effective Equations and Lorentz Invariance

A quantum state has infinitely many moments, which are all free to vary. The situation of states in quantum gravity is thus very different from a classical space-time such as Minkowski space. Also the fundamental dynamics of such states describing an atomic space-time is complicated and formulated in unfamiliar terms (examples can be found in [62–68]. The solution space currently is only poorly understood – too poorly to fully address important questions such as those about the status of classical symmetries. Lorentz symmetries are not manifest, even in the classical canonical formulation used to set up loop quantum gravity. It is then difficult to check if they are preserved after quantization, once all consistency conditions are implemented. (To make matters worse, it is not even fully known how to implement all the required consistency conditions in a complete manner.)

Fortunately, one can proceed at the level of effective equations and test whether effects such as those used in isotropic models are allowed in the much more constrained full setting. In particular, we saw that the discreteness, via the shift operators it allows to represent, implies higher order curvature corrections; and in quantizations of inverse powers of volume or the scale factor there are further deviations from classical behavior. This contributes correction terms to classical equations of motion, in addition to the coupling terms between expectation values and higher moments which arise in any interacting quantum system. If classical equations are corrected, classical symmetries may break. Once these terms are included in effective equations, symmetries can be tested much more easily than in the full quantum theory.

If the underlying quantum theory does not satisfy a classical symmetry, there will be terms in effective equations which explicitly show this breakdown. The converse is not generally true, however, and so it is not fully obvious to use effective equations as tests for symmetries. Effective equations for a Hamiltonian system (assumed here for the sake of the argument as being unconstrained) are obtained from the expectation value $\langle \hat{H} \rangle$ in a general state parameterized in terms of its moments [69, 70]. Symmetries of the fundamental dynamics given by $\hat{H}$ are to be tested, but effective equations also require a state, which enters dynamical equations via initial conditions for moments. Only in rare cases will the state be invariant under the classical symmetry, even if the whole theory respects the symmetry. In quantum field theory on Minkowski space, for instance, the vacuum is the only Poincaré invariant (pure) state. Low energy effective equations obtained by using the vacuum state to compute $\langle \hat{H} \rangle$ are thus manifestly covariant. But in other regimes where the vacuum state may not suffice, applying a symmetry of the quantum theory would change the state even if $\hat{H}$ is preserved, and therefore map one set of effective equations to another one.

A single set of effective equations, however, would not be invariant. This may be easy to see if the deviation from a vacuum state is brought in by switching on a background field, which then would have to be transformed under symmetries, too, implying that the set of effective equations has changed. In general, however, it can be difficult to discern which state properties are responsible for the breakdown of symmetries and how this can be reconciled in a symmetry-preserving full theory. Thus, effective equations which appear to break symmetries do not necessarily imply that they come from a full quantum theory which does not respects this symmetry.

Keeping this in mind, effective equations are a useful tool in this context because they allow one to test the anomaly issue more easily. For a constrained system, instead of the effective Hamiltonian $\langle\hat{H}\rangle$ we have effective constraints such as $\langle\hat{C}_I\rangle$ for a given set of classical constraints C_I (plus towers of higher moment constraints). For the effective constraints, one can now compute the Poisson brackets $\{\langle\hat{C}_I\rangle, \langle\hat{C}_J\rangle\}$ and see whether this forms a first class system. In general, effective constraints depend on all moments of the state, not just the expectation values. These variables thus have to be taken into account for the Poisson bracket, too, which can be computed through commutators, $\{\langle\cdot\rangle, \langle\cdot\rangle\} = \langle[\cdot,\cdot]\rangle/i\hbar$. In this way, one can systematically, and order by order in semiclassical or other expansions, check whether covariance is possible in the presence of quantum corrections, and how this restricts possible quantization choices.

As an example, one can consider quantum corrections to gravitational and electromagnetic waves. The respective Hamiltonians, weighted by the lapse function $N(x)$ to specify the spatial slicing, are

$$H_{\mathrm{G}} = \frac{1}{16\pi G}\int_\Sigma \mathrm{d}^3x N(x)\frac{E^c_j E^d_k}{\sqrt{|\det E|}}\left(\epsilon_i{}^{jk}F^i_{cd} - 2(1+\gamma^2)K^j_{[c}K^k_{d]}\right)$$

for the gravitational field (using as variables the densitized triad E^a_i, related to the spatial metric q_{ab} by $E^a_i E^b_i = q^{ab}\det q$, extrinsic curvature K^i_a, as well as the curvature F^i_{ab} of the Ashtekar connection [71] $A^i_a = \Gamma^i_a + \gamma K^i_a$, Γ^i_a being the spin connection and γ the Barbero–Immirzi parameter [72, 73]) and

$$H_{\mathrm{EM}} = \int_\Sigma \mathrm{d}^3x N(x)\left(\frac{2\pi}{\sqrt{q}}E^a E^b q_{ab} + \frac{\sqrt{q}}{16\pi}F_{ab}F_{cd}q^{ac}q^{bd}\right)$$

for the electromagnetic field, split into the electric field E^a and the magnetic field $B^a = \frac{1}{2}\epsilon^{abc}F_{bc}$. These classical expressions are subject to several corrections from a quantization. First, the gravitational Hamiltonian, depending on curvature, receives higher order corrections. Both Hamiltonians contain inverse powers of metric components, such that they receive additional corrections in loop quantum gravity. As always, there are finally coupling terms to fluctuations and higher moments, which would also involve the gravitational variables.

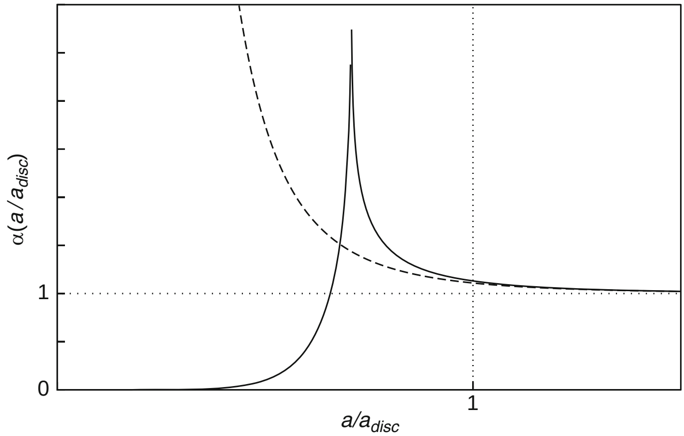


Fig. 8.4 Quantum correction function depending on a classical scale a relative to a scale a_{disc} determined by the underlying quantum state. For a/a_{disc} to the right of the peak position, the correction function is larger than one, which is the regime of perturbative calculations (Dashed line: asymptotic power law approximation.)

The simplest of those cases, and an especially instructive one, is that of corrections to the metric inverse. Explicit forms of corrections can be computed in some cases [74], but here it suffices to include correction functions as factors which differ from one but are otherwise unrestricted except that we will use the fact that corrections in perturbative regimes turn out to be larger than one, as can be seen from Fig. 8.4 (see also [75]).

Thus, we obtain corrected Hamiltonians

$$H_{\mathrm{G}}^{\mathrm{phen}} = \frac{1}{16\pi G}\int_{\Sigma} \mathrm{d}^3x N(x)\alpha(E_i^a)\frac{E_j^c E_k^d}{\sqrt{|\det E|}}\left(\epsilon_i{}^{jk}F_{cd}^i - 2(1+\gamma^2)K_{[c}^j K_{d]}^k\right)$$

and

$$H_{\mathrm{EM}}^{\mathrm{phen}} = \int_{\Sigma} \mathrm{d}^3x N(x)\left(\alpha_{\mathrm{EM}}(q_{cd})\frac{2\pi}{\sqrt{q}}E^a E^b q_{ab} + \beta_{\mathrm{EM}}(q_{cd})\frac{\sqrt{q}}{16\pi}F_{ab}F_{cd}q^{ac}q^{bd}\right)$$

with new, metric-dependent factors α, α_{EM} and β_{EM} which are all larger than one. (A superscript "phen" is now used rather than "eff" because these equations are not truly effective: only one type of corrections is included without checking the possible relevance of others.) Note that this is more general than simply multiplying all Hamiltonians with a factor which could be absorbed in the lapse function N, amounting only to a redefinition of time. In cases where $\alpha_{\mathrm{EM}} \neq \beta_{\mathrm{EM}}$ one could try to make them equal by a duality transformation of the electromagnetic field. However, this would have further complications for the canonical structure since α_{EM} and β_{EM} are triad dependent. We will further discuss this issue below.

The simplest property to analyze is the form of dispersion relations which can be derived from a plane wave ansatz. Doing so for the two fields results in linearized equations of motion [5]

$$\frac{1}{2}\left[\frac{1}{\alpha}\ddot{h}_a^i + 2\frac{\dot{a}}{a}\left(1 - \frac{2a\mathrm{d}\alpha/\mathrm{d}a}{\alpha}\right)\dot{h}_a^i - \alpha\nabla^2 h_a^i\right] = 8\pi G\Pi_a^i \tag{8.25}$$

for tensor modes h_a^i in a background Friedmann–Robertson–Walker space-time with source term Π_a^i. For electromagnetic waves, we obtain

$$\left(\alpha_{\mathrm{EM}}^{-1}\dot{A}_a\right)^{\cdot} - \beta_{\mathrm{EM}}\nabla^2 A_a = 0\,.$$

Thus, the quantum corrections affect the dispersion relations which, in a nearly flat space-time, become

$$\omega^2 = \alpha^2 k^2 \tag{8.26}$$

for gravitational waves and

$$\omega^2 = \alpha_{\mathrm{EM}}\beta_{\mathrm{EM}}k^2 \tag{8.27}$$

for electromagnetic waves.

This seems problematic because we have $\alpha > 1$, and similarly for the electromagnetic corrections. Thus, the speed of gravitational waves looks super-luminal, and it appears that the type of quantum correction used would break Lorentz invariance. At this point we again mention that one cannot simply repair this by absorbing α into the lapse function. This would mean that time proceeds differently due to the correction, and proper time derivatives of each field would be accompanied by an additional α^{-1} compared to classical equations. In this case, the α-factors in the source-free case would simply cancel in (8.25), and proper time equations of motion would have no corrections in their dispersion relations. However, this way out of Lorentz violations is not possible. Quantum corrections affect the Hamiltonian and thus equations of motion, but they do not change the space-time metric which in its canonical form

$$\mathrm{d}s^2 = -N^2\mathrm{d}t^2 + q_{ab}(\mathrm{d}x^a + N^a\mathrm{d}t)(\mathrm{d}x^b + N^b\mathrm{d}t)$$

dictates proper time as $\mathrm{d}\tau = N\mathrm{d}t$. This form of the metric does not receive quantum corrections, and there is no α multiplying N here even if there is such a factor in the Hamiltonian. What does change is the dynamics of q_{ab}, and for tensor modes this is captured in our linearized equation. The measurement of proper time as a kinematical procedure is not affected by quantum gravity, and so the dot used above unambiguously refers to proper time. In no way can the quantum correction simply be absorbed in a redefinition of time.

What appears super-luminal from (8.26) is the velocity of gravitational waves in quantum gravity compared to the *classical* speed of light. But this does not necessarily imply Lorentz violations because electromagnetic waves, as per (8.27) receive the same kind of quantum corrections, and thus the speed of light itself changes.

In fact, from the linearized equation for the vector potential we have a corrected dispersion relation which implies that "light is super-luminal." This paradoxical statement stresses the fact that comparison to the classical speed of light is meaningless in this context; the only physical statement would refer to a comparison between the quantum corrected speed of gravitational waves and the quantum corrected speed of light. Both differ from the classical speed of light, but a priori by different factors α and $\sqrt{\alpha_{\mathrm{EM}}\beta_{\mathrm{EM}}}$, respectively. Since these factors correct different classical expressions, and would, even for the same expression, allow several quantization ambiguities, they cannot be related to each other based only on the construction process of quantum operators. However, there are still consistency conditions to be imposed: both Hamiltonians form the Hamiltonian constraint and thus feature in the constraint algebra. For an anomaly-free calculation, the corrected expressions need to retain the first class form of the constraint algebra. As mentioned, this can be computed at the effective level, and although the calculation is lengthy, it is manageable. The result is simple: anomaly-freedom requires $\alpha^2 = \alpha_{\mathrm{EM}}\beta_{\mathrm{EM}}$ and removes any super-luminal motion.

8.5 Conclusions

A discrete structure of space may initially seem in danger of breaking Lorentz symmetries, in which case Minkowski space as the local form of space-times would only be a low-energy phenomenon: an approximation in which symmetry would allow economical calculations and powerful constructions, but itself devoid of any fundamental role. Its status, however, can only be final once all consistency conditions have been implemented, which is far from complete in loop quantum gravity due to technical difficulties.

Effective calculations are more manageable but still give promising indications. Indeed, it turns out that the loop quantization is more respectful to Lorentz symmetries than its discrete picture of space may suggest. (See also [76] for independent arguments.) This is, in the end, not fully surprising since anomaly-freedom of the space-time diffeomorphism algebra is required. But it is encouraging that one can verify this effectively, without detailed calculations of operator algebras, and that non-trivial quantum corrections are in fact allowed. In combination with quantum cosmology, where those quantum corrections can be used to shed light on the fate of the big bang, a consistent picture can be developed where low-energy physics works out in agreement with observations but high-energy phenomena still manage to improve on the classically singular behavior. Such a full picture is not yet available, for the corrections analyzed for their implications on anomaly freedom apply to different regimes than needed for a big bang analysis. But with improved derivations of effective equations these issues can be addressed systematically.

Moreover, even in homogeneous models a more complete treatment of states across the big bang is required to address the question of how strongly the quantum aspects of space-time can develop when a classical singularity is traversed. This is also required to understand how reliably one can speak of a state of the universe before the big bang. Currently, even in the simplest models and under rather strong conditions for states no strong restrictions on the behavior of the pre-big bang state are known. But much mathematical control on the behavior does exist thanks to the availability of a solvable model [6], in which one is not only in a lucky position to find explicit wave functions, but has full control on dynamical coherent states and thus the spreading behavior. Based on this model, a systematic perturbation theory is being developed which will allow one to understand the generic behavior over a wide range of models. In general, while only classical aspects of space-time are insufficient in such cases, the more general concept of a *quantum extended space-time* serves to understand the evolution of the universe from before the big bang.

Acknowledgements This work was supported in part by NSF grant PHY0653127.

References

1. Bojowald, M.: Loop quantum cosmology. Living Rev. Relat. **8**, 11 (2005) gr-qc/0601085, http://relativity.livingreviews.org/Articles/lrr-2005-11/
2. Bojowald, M.: Loop quantum cosmology. In: Ashtekar, A. (ed.) 100 Years of Relativity – Space-Time Structure: Einstein and Beyond, pp. 382–414. World Scientific, Singapore (2005) gr-qc/0505057
3. Bojowald, M.: What happened before the big bang? Nat. Phys. **3**, 523–525 (2007)
4. Bojowald, M.: Harmonic cosmology: how much can we know about a universe before the big bang? Proc. Roy. Soc. A **464**, 2135–2150 (2008) arXiv:0710.4919
5. Bojowald, M., Hossain, G.: Quantum gravity corrections to gravitational wave dispersion. Phys. Rev. D **77**, 023508 (2008) arXiv:0709.2365
6. Bojowald, M.: Large scale effective theory for cosmological bounces. Phys. Rev. D **75**, 081301(R) (2007) gr-qc/0608100
7. Weiss, N.: Constraints on Hamiltonian lattice formulations of field theories in an expanding universe. Phys. Rev. D **32**, 3228–3232 (1985)
8. Unruh, W.: Time, gravity, and quantum mechanics. In: Savitt, S.F. (ed.), Time's arrows today, pp. 23–94. gr-qc/9312027
9. Jacobson, T.: Trans-Planckian redshifts and the substance of the space-time river. hep-th/0001085
10. Doldán, R., Gambini, R., Mora, P.: Quantum mechanics for totally constrained dynamical systems and evolving Hilbert spaces. Int. J. Theor. Phys. **35**, 2057 (1996) hep-th/9404169
11. Bojowald, M.: The dark side of a patchwork universe. Gen. Rel. Grav. **40**, 639–660 (2008) arXiv:0705.4398
12. Bojowald, M.: Isotropic Loop Quantum Cosmology. Class. Quantum Grav. **19**, 2717–2741 (2002) gr-qc/0202077
13. DeWitt, B.S.: Quantum theory of gravity. I. The canonical theory. Phys. Rev. **160**, 1113–1148 (1967)
14. Wiltshire, D.L.: An introduction to quantum cosmology. In: Robson, B., Visvanathan, N., Woolcock, W.S. (eds.), Cosmology: The Physics of the Universe, 473–531. World Scientific, Singapore (1996) gr-qc/0101003
15. Bojowald, M.: Loop quantum cosmology and inhomogeneities. Gen. Rel. Grav. **38**, 1771–1795 (2006) gr-qc/0609034
16. Ashtekar, A., Pawlowski, T., Singh, P.: Quantum nature of the Big Bang: improved dynamics. Phys. Rev. D **74** 084003 (2006) gr-qc/0607039

17. Bojowald, M., Cartin, D., Khanna, G.: Lattice refining loop quantum cosmology, anisotropic models and stability. Phys. Rev. D **76**, 064018 (2007) arXiv:0704.1137
18. Nelson, W., Sakellariadou, M.: Lattice refining LQC and the matter Hamiltonian. Phys. Rev. D **76**, 104003 (2007) arXiv:0707.0588
19. Nelson, W., Sakellariadou, M.: Lattice refining loop quantum cosmology and inflation. Phys. Rev. D **76**, 044015 (2007) arXiv:0706.0179
20. Bojowald, M., Hossain, G.: Cosmological vector modes and quantum gravity effects. Class. Quantum Grav. **24**, 4801–4816 (2007) arXiv:0709.0872
21. Bojowald, M.: Loop quantum cosmology IV: discrete time evolution. Class. Quantum Grav. **18**, 1071–1088 (2001) gr-qc/0008053
22. Bojowald, M.: Inverse Scale Factor in Isotropic Quantum Geometry. Phys. Rev. D **64**, 084018 (2001) gr-qc/0105067
23. Thiemann, T.: QSD V: Quantum gravity as the natural regulator of matter quantum field theories. Class. Quantum Grav. **15**, 1281–1314 (1998) gr-qc/9705019
24. Singh, P.: Loop cosmological dynamics and dualities with Randall-Sundrum braneworlds. Phys. Rev. D **73**, 063508 (2006) gr-qc/0603043
25. Bojowald, M.: Dynamical coherent states and physical solutions of quantum cosmological bounces. Phys. Rev. D **75**, 123512 (2007) gr-qc/0703144
26. Corichi, A., Singh, P.: Quantum bounce and cosmic recall, Phys. Rev. Lett. **100**, 161302 (2008) arXiv:0710.4543
27. Bojowald, M., Hernández, H., Skirzewski, A.: Effective equations for isotropic quantum cosmology including matter. Phys. Rev. D **76**, 063511 (2007) arXiv:0706.1057
28. Date, G.: Absence of the Kasner singularity in the effective dynamics from loop quantum cosmology. Phys. Rev. D **71**, 127502 (2005) gr-qc/0505002
29. Modesto, L.: Black hole interior from loop quantum gravity. Adv. High Energy Phys. **2008**, 459290 (2008) gr-qc/0611043
30. Modesto, L.: Evaporating loop quantum black hole. gr-qc/0612084
31. Chiou, D.-W.: Effective dynamics for the cosmological bounces in Bianchi type I loop quantum cosmology. arXiv:gr-qc/0703010
32. Chiou, D.-W., Vandersloot, K.: The behavior of non-linear anisotropies in bouncing Bianchi I models of loop quantum cosmology. Phys. Rev. D **76**, 084015 (2007) arXiv:0707.2548
33. Chiou, D.-W.: Effective dynamics, big bounces and scaling symmetry in Bianchi type I loop quantum cosmology. Phys. Rev. D **76**, 124037 (2007) arXiv:0710.0416
34. Böhmer, C.B., Vandersloot, K.: Loop quantum dynamics of the Schwarzschild interior. Phys. Rev. D **76**, 104030 (2007) arXiv:0709.2129
35. Campiglia, M., Gambini, R., Pullin, J.: Loop quantization of spherically symmetric midi-superspaces: the interior problem. AIP Conf. Proc. **977**, 52–63 (2008) arXiv:0712.0817
36. Bojowald, M.: Singularities and Quantum Gravity. AIP Conf. Proc. **910**, 294–333 (2007) gr-qc/0702144, Proceedings of the XIIth Brazilian School on Cosmology and Gravitation
37. Ashtekar, A., Bojowald, M.: Quantum geometry and the Schwarzschild singularity. Class. Quantum Grav. **23**, 391–411 (2006) gr-qc/0509075
38. Modesto, L.: The Kantowski-Sachs space-time in loop quantum gravity. Int. J. Theor. Phys. **45**, 2235–2246 (2006) gr-qc/0411032
39. Bojowald, M.: Absence of a singularity in loop quantum cosmology. Phys. Rev. Lett. **86**, 5227–5230 (2001) gr-qc/0102069
40. Bojowald, M.: Homogeneous loop quantum cosmology. Class. Quantum Grav. **20**, 2595–2615 (2003) gr-qc/0303073
41. Bojowald, M., Date, G., Vandersloot, K.: Homogeneous loop quantum cosmology: The role of the spin connection. Class. Quantum Grav. **21**, 1253–1278 (2004) gr-qc/0311004
42. Bojowald, M.: Spherically symmetric quantum geometry: states and basic operators. Class. Quantum Grav. **21**, 3733–3753 (2004) gr-qc/0407017
43. Bojowald, M., Swiderski, R.: Spherically symmetric quantum geometry: Hamiltonian constraint. Class. Quantum Grav. **23**, 2129–2154 (2006) gr-qc/0511108
44. Banerjee, K., Date, G.: Loop quantization of polarized Gowdy model on T^3: classical theory. Class. Quantum Grav. **25**, 105014 (2008) arXiv:0712.0683

45. Banerjee, K., Date, G.: Loop quantization of polarized Gowdy model on T^3: quantum theory. Class. Quantum Grav. **25**, 145004 (2008) arXiv:0712.0687
46. Ashtekar, A., Bojowald, M.: Black hole evaporation: a paradigm. Class. Quantum Grav. **22**, 3349–3362 (2005) gr-qc/0504029
47. Bojowald, M., Goswami, R., Maartens, R., Singh, P.: A black hole mass threshold from non-singular quantum gravitational collapse. Phys. Rev. Lett. **95**, 091302 (2005) gr-qc/0503041
48. Bojowald, M.: Quantum Riemannian Geometry and Black Holes. Nova Science (2006) gr-qc/0602100
49. Bojowald, M.: Non-singular black holes and degrees of freedom in quantum gravity, Phys. Rev. Lett. **95**, 061301 (2005) gr-qc/0506128
50. Husain, V., Winkler, O.: Quantum resolution of black hole singularities. Class. Quantum Grav. **22**, L127–L133 (2005) gr-qc/0410125
51. Bojowald, M., Swiderski, R.: Spherically Symmetric Quantum Horizons. Phys. Rev. D **71**, 081501(R) (2005) gr-qc/0410147
52. Husain, V., Winkler, O.: Quantum black holes from null expansion operators. Class. Quantum Grav. **22**, L135–L141 (2005) gr-qc/0412039
53. Husain, V., Winkler, O.: How red is a quantum black hole? Int. J. Mod. Phys. D **14**, 2233–2238 (2005) gr-qc/0505153
54. Husain, V., Winkler, O.: Quantum Hamiltonian for gravitational collapse. Phys. Rev. D **73**, 124007 (2006) gr-qc/0601082
55. Roman, T.A., Bergmann, P.G.: Stellar collapse without singularities? Phys. Rev. D **28**, 1265–1277 (1983)
56. Hayward, S.A.: Formation and evaporation of non-singular black holes. Phys. Rev. Lett. **96**, 031103 (2006) gr-qc/0506126
57. Cartin, D., Khanna, G.: Wave functions for the Schwarschild black hole interior. Phys. Rev. D **73**, 104009 (2006) gr-qc/0602025
58. Bojowald, M.: Dynamical Initial Conditions in Quantum Cosmology. Phys. Rev. Lett. **87**, 121301 (2001) gr-qc/0104072
59. Bojowald, M.: Initial Conditions for a Universe. Gen. Rel. Grav. **35**, 1877–1883 (2003) gr-qc/0305069
60. Rosen, J., Jung, J.-H., Khanna, G.: Instabilities in numerical loop quantum cosmology. Class. Quantum Grav. **23**, 7075–7084 (2006) gr-qc/0607044
61. Sabharwal, S., Khanna, G.: Numerical solutions to lattice-refined models in loop quantum cosmology. Class. Quantum Grav. **25**, 085009 (2008) arXiv:0711.2086
62. Loll, R.: Discrete approaches to quantum gravity in four dimensions, Living Rev. Rel. **1**, 13 (1998) gr-qc/9805049. http://www.livingreviews.org/lrr-1998-13
63. Thiemann, T.: Quantum Spin Dynamics (QSD). Class. Quantum Grav. **15**, 839–873 (1998) gr-qc/9606089
64. Thiemann, T.: The phoenix project: master constraint programme for loop quantum gravity. gr-qc/0305080
65. Perez, A.: Spin foam models for quantum gravity. Class. Quantum Grav. **20**, R43 (2003) gr-qc/0301113
66. Giesel, K., Thiemann, T.: Algebraic quantum Gravity (AQG) I. Conceptual setup. Class. Quantum Grav. **24**, 2465–2497 (2007) gr-qc/0607099
67. Konopka, T., Markopoulou, F., Smolin, L.: Quantum graphity. hep-th/0611197
68. Markopoulou, F., Smolin, L.: Disordered locality in loop quantum gravity states. Class. Quant. Grav. **24**, 3813–3824 (2007) gr-qc/0702044
69. Bojowald, M., Skirzewski, A.: Effective equations of motion for quantum systems. Rev. Math. Phys. **18**, 713–745 (2006) math-ph/0511043
70. Bojowald, M., Skirzewski, A.: Quantum gravity and higher curvature actions. Int. J. Geom. Meth. Mod. Phys. **4**, 25–52 (2007) hep-th/0606232. Proceedings of "Current Mathematical Topics in Gravitation and Cosmology" (42nd Karpacz Winter School of Theoretical Physics), Ed. Borowiec, A., Francaviglia, M.
71. Ashtekar, A.: New Hamiltonian formulation of general relativity. Phys. Rev. D **36**, 1587–1602 (1987)

72. Fernando J., Barbero G.: Real Ashtekar variables for Lorentzian signature space-times. Phys. Rev. D **51**, 5507–5510 (1995) gr-qc/9410014
73. Immirzi, G.: Real and complex connections for canonical gravity. Class. Quantum Grav. **14**, L177–L181 (1997)
74. Bojowald, M., Hernández, H., Kagan, M., Skirzewski, A.: Effective constraints of loop quantum gravity. Phys. Rev. D **75**, 064022 (2007) gr-qc/0611112
75. Bojowald, M., Das, R., Scherrer, R.: Dirac fields in loop quantum gravity and big bang nucleosynthesis. Phys. Rev. D **77**, 084003 (2008) arXiv:0710.5734
76. Rovelli, C., Speziale, S.: Reconcile Planck-scale discreteness and the Lorentz-Fitzgerald contraction. Phys. Rev. D **67**, 064019 (2003) gr-qc/0205108

Part III
Conceptual and Philosophical Issues of Minkowski Spacetime

Chapter 9
The Adolescence of Relativity: Einstein, Minkowski, and the Philosophy of Space and Time

Dennis Dieks

Abstract An often repeated account of the genesis of special relativity tells us that relativity theory was to a considerable extent the fruit of an operationalist philosophy of science. Indeed, Einstein's 1905 paper stresses the importance of rods and clocks for giving concrete physical content to spatial and temporal notions. I argue, however, that it would be a mistake to read too much into this. Einstein's operationalist remarks should be seen as serving rhetoric purposes rather than as attempts to promulgate a particular philosophical position – in fact, Einstein never came close to operationalism in any of his philosophical writings. By focussing on what could actually be measured with rods and clocks Einstein shed doubt on the empirical status of a number of pre-relativistic concepts, with the intention to persuade his readers that the applicability of these concepts was not obvious. This rhetoric manoeuvre has not always been rightly appreciated in the philosophy of physics. Thus, the influence of operationalist misinterpretations, according to which associated operations strictly *define* what a concept means, can still be felt in present-day discussions about the conventionality of simultaneity.

The standard story continues by pointing out that Minkowski in 1908 supplanted Einstein's approach with a realist spacetime account that has no room for a foundational role of rods and clocks: relativity theory became a description of a four-dimensional "absolute world." As it turns out, however, it is not at all clear that Minkowski was proposing a substantivalist position with respect to spacetime. On the contrary, it seems that from a philosophical point of view Minkowski's general position was not very unlike the one in the back of Einstein's mind. However, in Minkowski's formulation of special relativity it becomes more explicit that the content of spatiotemporal concepts relates to considerations about the form of physical laws. If accepted, this position has important consequences for the discussion about the conventionality of simultaneity.

Keywords Special relativity · Conventionalism · Operationalism · Simultaneity · Einstein · Minkowski

D. Dieks (✉)
History and Foundations of Science, Utrecht University,
P.O. Box 80.010, NL 3508 TA Utrecht, The Netherlands
e-mail: d.dieks@uu.nl

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_9,

9.1 Introduction

At the end of the introductory section of his "On the electrodynamics of moving bodies" Einstein [8, p. 892; 24, p. 277] famously declares: "The theory to be developed is based – like all electrodynamics – on the kinematics of the rigid body, since the assertions of any such theory have to do with the relationships between rigid bodies (systems of co-ordinates), clocks, and electromagnetic processes. Insufficient consideration of this circumstance lies at the root of the difficulties which the electrodynamics of moving bodies at present encounters" (English translation from [17, p. 38]). When Einstein subsequently starts discussing the notion of time he elaborates on the same point and warns us that a purely theoretical, mathematical, description "has no physical meaning unless we are quite clear as to what we understand by 'time'." He goes on by explaining that we need to provide our concepts with concrete physical content and that for the case of time at one spatial position the sought *definition* (Einstein's term) can simply be given as "the position of the hands of my watch" (situated at the position in question). Time thus defined is a purely local concept, however, so that we need a further definition in order to compare times at different positions. For this reason Einstein famously engages in a discussion of simultaneity. He briefly considers the possibility of assigning to distant events the time indicated by one fixed clock at the moment a light signal from the events reaches this clock, but rejects this possibility because the time thus assigned would depend on the location of the standard clock (which would have as a consequence that physical laws would become position-dependent too). A much better idea is to work with synchronized clocks without any hierarchical ordering between them. This then finally leads to the introduction of Einstein's famous procedure for synchronizing clocks: synchronicity is *by definition* achieved when all clocks are set such that the velocity of light, measured with their help, becomes the same in all directions. Now the characterization of time in a frame of reference has become complete: "the 'time' of an event is the indication which is given simultaneously with the event by a stationary clock located at the place of the event, where this clock should be synchronous for all time determinations with a specified stationary clock" [24, p. 279; 17, p. 40].

These passages, and others in the 1905 paper, appear to put forward an undeniably operationalist conception of spatial and temporal notions. Coordinates are identified with notches in rigid material axes, distances *are* what is measured by rigid measuring rods, and time *is* what is indicated by the hands of synchronized clocks. This operationalist flavour becomes even stronger because Einstein repeatedly uses the term "definition" in his analysis: time is *defined* via operations with clocks, and thus apparently has no other meaning than what results from these operations. Because definitions, as the term is used in its natural scientific habitat – logic and mathematics – are the results of our free decisions and cannot be true or false, this suggests that Einstein is telling us here that spatial and temporal notions, among them "simultaneity", are purely conventional in character. So a twofold philosophical message seems to be implied: first, fundamental physical concepts must be defined via concrete physical operations in order to be meaningful at all and second, these definitions have the status of conventions.

Einstein's statements in these pages have had an enormous influence in twentieth century philosophy of science. Among logical positivists it was one of the motivations for developing the doctrine of "coordinative definitions," according to which physical concepts (like "time") should be coordinated to concrete physical things and procedures. Schlick [22], and in his footsteps [21], emphasized that this coordinatization is fundamentally conventional in character; Reichenbach elaborated this idea in famous detail in his analysis of simultaneity. Percy Bridgman, the founder of operationalism, took his inspiration from Einstein's analysis as well. In his contribution to *Albert Einstein: Philosopher-Scientist* [23], Bridgman [1] wrote: "Let us examine what Einstein did in his special theory. In the first place, he recognized that the meaning of a term is to be sought in the operations employed in making application of the term. If the term is one which is applicable to concrete physical situations, as "length" or "simultaneity", then the meaning is to be sought in the operations by which the length of concrete physical objects is determined, or in the operations by which one determines whether two concrete physical events are simultaneous or not." Bridgman went on to complain that in his General Theory of Relativity Einstein seemed to have forgotten some of his own methodological lessons.

Bridgman must have been disillusioned by the reply Einstein gave to his admonitions. In his "Remarks to the Essays Appearing in this Collective Volume" [23], Einstein squarely rejected operationalism, both in the context of special and general relativity. That Einstein here opposed operationalism in such strong terms is remarkable: did Einstein change his beliefs, or remained his convictions more or less the same and are there ways of reading his 1905 statements other than those proposed by the logical positivists and operationalists? Of course, one must be careful in interpreting the documentary material here: what Einstein wrote later in his career need not at all faithfully reflect his attitudes as a young scientist. However, after reviewing the evidence, we shall indeed conclude in this paper that another, non-operationalist, reading of Einstein's early work is appropriate. But let us first return to what perhaps may be called "the standard account".

The usual story about the genesis of special relativity theory says that after Einstein's operationalist introduction of the theory, Minkowski in 1908 proposed a very different view. Minkowski interpreted special relativity as a geometrical description of a four-dimensional spacetime manifold, which he called "the Absolute World". According to this new point of view relativity theory does not depend on the coordinatization of events with the help of rods and clocks: it is a theory about an independent spacetime manifold that possesses an inbuilt geometrical structure and subsists even if there are no rods and clocks at all.

Although it would be an exaggeration to say that this characterization of Minkowski's work misses the mark altogether, I believe that both the emphasis on the difference between Minkowski and Einstein *qua* philosophical outlook, and the insistence that Minkowski posited the independent reality of a four-dimensional spacetime manifold, are misplaced. As I shall argue, Einstein no less than Minkowski thought of relativity as a theory about the general form of physical laws, without any special status for rods and clocks. Further, Einstein had misgivings about space and time as entities existing by themselves; but likewise Minkowski expressed doubts about empty spacetime and stated as his belief that special relativity is best seen as a theory about the relations between material systems. Although it is true that Einstein's presentation in which rods and clocks figure prominently is replaced by Minkowski with one in which fundamental particles and fields are basic, I believe that this does not signify a fundamental philosophical difference between Einstein and Minkowski.

Indeed, the picture that will emerge from the analysis given in this paper is that Minkowski did not give a completely new philosophical turn to relativity theory but rather completed, in a mathematically sophisticated and elegant way, the programme that Einstein had in mind. This conclusion has consequences for the status of the spacetime concepts that occur in the Einstein-Minkowski theory. In particular, new light is shed on the status of relativistic simultaneity. This is the final issue that we shall address. We shall argue that from a correctly understood Einsteinian–Minkowskian viewpoint simultaneity in special relativity is not more conventional than other fundamental physical concepts.

9.2 Einstein and the Definition of Space and Time

The 1905 paper is not the only place where Einstein expresses himself in a way that suggests operationalist sympathies. It is striking that even much later, in the Autobiographical Notes – the very same volume we referred to above [23] – we find Einstein reminiscing about his discovery of special relativity with the following words: "One had to understand clearly what the spatial co-ordinates and the temporal duration of events meant in physics. The physical interpretation of the spatial co-ordinates presupposed a fixed body of reference, which, moreover, had to be in a more or less definite state of motion (inertial system). In a given inertial system the co-ordinates meant the results of certain measurements with rigid (stationary) rods. If, then, one tries to interpret the time of an event analogously, one needs a means for the measurement of the difference in time. A clock at rest relative to the system of inertia defines a local time. The local times of all space points taken together are the 'time' which belongs to the selected system of inertia, if a means is given to 'set' these clocks relative to each other" [23, p. 55]. Historians of science often warn us not to rely too much on (much) later accounts scientists give of the way in which they made their discoveries: later experiences may very well have coloured and distorted their memories. But here we find an almost verbatim repetition of the relevant passages from the 1905 paper itself, including the use of the term *define*, and with the explanation that space and time coordinates *mean* what is indicated by rods and clocks; and all this without any accompanying comment that might indicate that Einstein in the nineteen-forties deemed some kind of qualification of his 1905 statements necessary. So we may safely assume that Einstein is here expressing the same view as the one he put forward in his original relativity paper. This is striking, most of all because elsewhere in these same autobiographical notes, and also in Einstein's "Replies" in the same volume [23] we find an explicit and strong rejection of operationalism as a viable philosophy of science.

As we already mentioned, Einstein's resented Bridgman's characterization of special relativity as a fountainhead of operationalism. The essential part of his Reply reads: "In order to be able to consider a logical system as physical theory it is not necessary to demand that all of its assertions can be independently interpreted and 'tested' 'operationally'; *de facto* this has never been achieved by any theory and can not at all be achieved. In order to be able to consider a theory as a *physical* theory it is only necessary that it implies empirically testable assertions in general" [23, p. 679]. Einstein made the same point in greater detail in his Reply to Reichenbach. In his contribution to the Einstein Volume, Reichenbach had contended that the philosophical lesson to be drawn from relativity theory was that basic physical concepts must be given meaning by means of "co-ordinative definitions": it is only the "co-ordination" of a concrete physical object or process to the concepts in question that bestows physical significance on them. "For instance," Reichenbach wrote [23, p. 295], "the concept "equal length" is defined by reference to a physical object, a solid rod, whose transport lays down equal distances. The concept "simultaneous" is defined by the use of light-rays which move over equal distances. The definitions of the theory of relativity are all of this type; they are co-ordinative definitions." Reichenbach continued by explaining that this definitional character of basic physical concepts implies that they are *arbitrary*. "Definitions are arbitrary; and it is a consequence of the definitional character of fundamental concepts that with the change of the definitions various descriptional systems arise. Thus the definitional character of the fundamental concepts leads to a plurality of equivalent descriptions. All these descriptions represent different languages saying the same thing; equivalent descriptions, therefore, express the same physical content." In his response Einstein famously staged a dialogue between Reichenbach and Poincaré, later in the dialogue replaced by a "non-positivist"; Einstein himself clearly being on the side of the non-positivist camp. Against the idea of coordinative definitions Einstein levelled the objection that any concrete physical object is subject to deforming forces, and can therefore not be used to *define* concepts. We need a *theory* of these deforming influences in order to be able to correct for them, and such a theory already uses a notion of length. Therefore, we must know what "length" is prior to the determination of the undisturbed length of any measuring rod. From this Einstein concludes that a concept like "equality of length" cannot be defined by reference to concrete objects at all; such concepts "are only indispensable within the framework of the logical structure of the theory, and the theory validates itself only in its entirety [23, p. 678]."

These remarks are in accordance with Einstein's often-expressed conviction that scientific theories and laws cannot be *derived* from experience but must prove their value when, once formulated as "free creations of the human mind", they are confronted as a whole with experience. With respect to space and time, we find this attitude clearly present in the lecture *Geometry and Experience* ([14, pp. 232–246];

German original *Geometrie und Erfahrung* [12]). In *Geometry and Experience* Einstein writes: "The idea of the measuring rod and the idea of the clock in the theory of relativity do not find their exact correspondence in the real world. It is also clear that the solid body and the clock do not in the conceptual edifice of physics play the part of irreducible elements, but that of composite structures, which must not play any independent part in theoretical physics." One might wonder how this statement, made relatively soon after the discovery of relativity theory, can be squared with the role assigned to rods and clocks in the 1905 paper. Einstein is quick to answer this question. In the same 1921 lecture he continues: "It is my conviction that in the present stage of development of theoretical physics these concepts (i.e., rods and clocks) must still be employed as independent concepts; for we are still far from possessing such certain knowledge of the theoretical principles of atomic structure as to be able to construct solid bodies and clocks theoretically from elementary concepts." Einstein goes on by explaining that the problem of deforming forces need not be prohibitive in practice: by comparing different solid bodies, of different constitution, we may obtain information about the order of magnitude of the deformations and we can then make appropriate corrections that suffice for practical purposes.

In a short, and not very well known, contribution to the German literary journal *die neue Rundschau*, Einstein [13] attempted to explain the situation to a general audience. To answer the question whether Euclidean geometry or some other geometry applies to the physical world, Einstein tells us, we have to choose between two possible points of view. Either we assume that geometrical concepts correspond, in an approximate fashion, to concrete physical objects — this is the attitude of the working physicist, and without it the creation of relativity theory would have been impossible. Or one assumes from the beginning that geometry by itself is not about real objects, but that only the *combination* of geometry and physics makes contact with physical reality. The latter point of view is probably, Einstein says, the one that is best for a systematic presentation of an already fully elaborated physics, of which we know the laws. In this case, the answer to the question of which geometry pertains to the physical world depends on how 'simple' the associated physics becomes when we choose the geometry in question.

The first point that consistently emerges from these statements is that the unit of length may only be supposed to be realized by a "theoretical" object, an *ideal* rod, which can merely be approximated by concrete objects. Although there is no reference to deforming forces and theoretical approximations in the 1905 paper, we may safely assume that even at this early stage Einstein, having received part of his physics training in the laboratory, saw the situation in this light – however, there was evidently little motivation for him to dwell on these distracting epistemological issues in the context of the introduction of his radically new physical ideas. That actual concrete objects do not fully represent theoretical space-time standards is an almost obvious thing, from a physical point of view. Indeed, in his early review article on special relativity, *Über das Relativitätsprinzip und die aus demselben gezogen Folgerungen*, written very soon after the 1905 paper, Einstein [9] repeated that for the assignment of spatial coordinates rigid rods are needed; but he added a footnote

[24, p. 437] saying that instead of referring to "rigid" rods we could as well speak of solid bodies "not subject to deforming forces", clearly indicating that he was aware of the complications.

A second point that stands out is that Einstein did not think that rods and clocks have a truly foundational role to play. It is only because in everyday situations we are accustomed to thinking of rods and clocks for determining space and time coordinates, and are not able to directly describe these devices in terms of fundamental physical theory, that it is expedient to introduce them to fix what we are talking about. This is a practical decision, made for the time being; "with the obligation, however, of eliminating it at a later stage of the theory" (Einstein in [23, p. 59]). Thus, Einstein says that as soon as a direct characterisation via fundamental physical theory becomes available, this treatment will have to replace the rods-and-clocks account on the level of foundational considerations. This pragmatic attitude with respect to rods and clocks is a far cry from the idea that these devices *define* length and time!

There are two documents from the first few years after 1905 in which Einstein makes more than casual remarks on the status of simultaneity. In a 1910 paper published in French ([10]; [16], pp. 131–174) Einstein gives a general overview of special relativity, with particular attention to its epistemological foundations. As he wrote in a letter to Laub ([16], p. 175) this paper "comprises a rather general discussion of the epistemological foundations of the theory of relativity, no new views whatsoever" ("eine ziemlich breite Ausführung der erkenntnistheoretischen Grundlagen der Relativitätstheorie, gar keine neue Überlegungen"). In this article we find an extensive and interesting discussion of the relativistic conception of time. The discussion ([16], pp. 146–147) starts with the observation that we can *measure* time with the help of clocks ("Pour mesurer le temps nous nous servirons d'horloges"), continues with the assertion that we are obliged, by the principle of sufficient ground, to admit that subsequent periods of a clock take equally long periods of time ("... que nous soyons obligés d'admettre – en vertu du principe de raison suffisante – que...."), but then suddenly shifts to definition-terminology: the number of periods indicated by a clock *defines* the lapse of time, Einstein concludes. He goes on by stating that this definition of "local time" is not enough; we need to say something additional about simultaneity ("La définition est alors insuffisante: il faut la compléter.") He explains this with an extensive exposition about the importance of setting clocks with respect to each other for the description of processes that are not restricted to one spatial position, and ends with a remarkable explanation of the synchronisation procedure.

"Let us first make available a means for sending signals from A to B and vice versa. This means must be such that we have no reason whatsoever to believe that the transmission phenomena in the direction AB differ in any respect from those in the direction BA. In this case, it is evident that there is only one way to set the clocks in B and A so that the signal from A to B takes the same time – measured with the mentioned clocks – as the one going from B to A." ("Donnons-nous d'abord un moyen pour envoyer des signaux soit de A en B, soit de B en A. Ce moyen doit être tel que nous n'ayons aucune raison pour croire que les phénomènes de transmission des signaux dans le sens AB diffèrent en quelque chose des phénomènes de transmission des signaux dans le sens BA. Dans ce cas, il est manifeste qu'il n'y a qu'une seule manière de régler l'horloge de B sur celle de A de façon que le signal allant de A en B prenne autant de temps – mesuré à l'aide des dites horloges – que celui allant de B en A" ([16], p. 149)). After this justification, the standard ("$\varepsilon = 1/2$") formula appears. But Einstein is not ready yet: he continues by pointing out that in principle we could use *any* signal for this procedure as long as we can be sure about the equal signal speeds in the two directions. "But we will give preference to light signals in vacuum, because as the synchronization requires the equivalence of the to and fro ways, we shall have this equivalence by definition since by virtue of the principle of the constancy of the speed of light, light in vacuum always propagates with the speed c. We therefore shall have to set our clocks such that the time for a light signal to go from A to B will equal that needed by a similar signal to go from B to A." ("Cependant, ... nous donnerons notre préférence à ceux où l'on fait usage de rayons lumineux se propageant dans le vide, car, le réglage exigeant l'équivalence du chemin d'aller avec celui du retour, nous aurons alors cette équivalence par définition, puisque, en vertu du principe de la constance de la vitesse de la lumière, la lumière dans le vide se propage toujours avec la vitesse c. Nous devrons donc régler nos horloges de façon que le temps employé par un signal lumineux pour aller de A en B soit égal à celui employé par un même signal allant de B en A" ([16], p. 150)).

It is true that the word "definition" occurs in this exposition, but the text makes it clear that no arbitrary stipulation is meant. Rather, the idea expressed by "defining" simultaneity via sending light signals to and fro is that we can be sure about the validity of this synchronization procedure, given the light principle. Similarly, in the case of the comparison of successive periods of a periodic process (a clock) we saw Einstein invoking the principle of sufficient reason to justify the "definition" of the equality of these intervals. Evidently, Einstein had no qualms in letting the justification of such "definitions" depend on prior theoretical principles, even if these very principles need the "definitions" in question if we want to *test* them. The light principle, e.g., can only be empirically verified if we know how to measure the speed of light; and for this we need synchronized clocks. That Einstein was fully aware of this complication is already clear in the 1905 paper, but we find a more explicit discussion of this point and its relevance in the text of a lecture delivered by him in 1911 ([11]; [16], pp. 425–438).

This lecture, presented at a meeting of the *Naturforschende Gesellschaft Zürich*, gives a historical introduction to special relativity from the point of view of an experimentalist: centre stage is taken by the question of what can actually be measured and what the experimental support for the relativistic principles is. In connection with time this question becomes: How can we characterize time in such a way that we can actually measure it? The complication is in the determination of simultaneity, as Einstein explains ([16], p. 431): to synchronize clocks we need to know the speed of the signals by means of which we set these clocks with respect to each other, but this speed can in turn only be measured if we already have synchronized clocks at our disposal. This vicious measurement-circle makes it possible for us to make certain stipulations with regard to the speeds of signals, in particular the speed of light; and we use this to lay down that the speed of light in vacuum from A to B equals the speed from B to A. ("Wenn es nun aber ohne willkürliche Festsetzung prinzipiell ausgeschlossen ist, eine Geschwindigkeit, im speziellen die Geschwindigkeit des Lichts, zu messen, so sind wir berechtigt, bezüglich der Fortpflanzungsgeschwindigkeit des Lichtes noch willkürliche Festsetzungen zu machen. Wir setzen nun fest, dass die Fortpflanzungsgeschwindigkeit des Lichtes im Vacuum auf dem Wege von einem Punkt A nach einem Punkt B gleich gross sei wie die Fortplanzungsgeschwindigkeit eines Lichtstrahls von B nach A" ([16], 432).) With this, it becomes possible to unambiguously synchronize clocks: we have achieved a determination of time "from the standpoint of the measuring physicist" ("so haben wir eine Zeitbestimmung vom Standpunkt des messenden Physikers erlangt" ([16], 432)).

The text of this lecture is the one coming closest to the idea that relativistic simultaneity rests on an arbitrary definition; indeed, the word "arbitrary" ("willkürlich") occurs explicitly. But the context, the discussion of the possibilities of actual measurement, makes it clear that what is said here is only that simultaneity is not already determined, for the "measuring physicist", by the system of clocks at various places. This measuring physicist wants to have a concrete operational recipe for physical quantities, and for this something additional is needed. Once this lacuna is recognized, it is filled up immediately by Einstein by means of the light principle, without any discussion of possible alternatives. Seen this way, there is no conflict between the 1911 lecture text and the extensive analysis of the year before.

Putting all this evidence into one coherent whole, it is natural to conclude that Einstein's reference to "definitions", in his discussion of relativistic concepts, should not be taken as embodying a systematic operationalist or logical empiricist philosophy. There is a remarkable continuity and constancy in Einstein's utterances from the early twenties onwards, when he first explicitly addresses philosophical questions relating to space and time. In these philosophically inclined writings Einstein consistently rejects the project of defining concepts along the lines of operationalism or logical empiricism. The striking fact that Einstein uses the term "definition," in this very context of anti-operationalist reflections and much later than 1905, demonstrates that he did not realize the extent to which this term could excite philosophers and could give rise to misunderstandings. Actually, it is quite understandable that Einstein used the term "definition" in a way that did not fully accord with its use in logic or philosophy. Einstein's papers on special relativity, in which this term figures so prominently, are obviously *physics* papers, addressed to a physicist audience. In these papers Einstein was facing the task of convincing his readers that the spatiotemporal concepts of classical physics were not beyond discussion; what could be a better strategy to accomplish this aim than showing that actual measurements do not provide support for the applicability of these classical concepts? This strategy explains the emphasis on measurement procedures. In the rhetorical context it then is a natural step to reinforce the argument by speaking about the measurement procedures as "definitions" – a term that sounds stronger and more definitive than "measurement" or "determination".

Here it should be added that in spite of the fact that philosophically speaking the status of the meaning of spatial and temporal concepts is the same, Einstein in his 1905 paper only uses the term "definition" when he discusses the in his eyes problematical and to-be-changed concept of *time*. In the case of the spatial coordinates he simply speaks about "determining" or "measuring" the coordinates. Thus, in the beginning of Section 1, *Definition of Simultaneity*, Einstein writes [24, p. 277]: "Ruht ein materieller Punkt relativ zu diesem Koordinatensystem, so kann seine Lage relativ zu letzterem durch starre Maßstäbe unter Benutzung der Methoden der Euklidischen Geometrie *bestimmt* und in kartesischen Koordinaten ausgedrückt werden" (my emphasis). The translation by Perrett and Jeffery [17] renders this as: "If a material point is at rest relatively to this system of co-ordinates, its position can be *defined* relatively thereto by the employment of rigid standards of measurement and the methods of Euclidean geometry, and can be expressed in Cartesian co-ordinates" (my emphasis). In the context of the present discussion this translation is unfortunate. The German text does not speak about definitions: "Bestimmt" simply means "determined", and does not possess the mathematical-logical-philosophical connotations of "defined". Nevertheless, the translation cannot be called incorrect, since in physics the verb "to define" is frequently used in a loose manner, without its philosophical connotations. An example of this can be found in the 1905 special relativity paper itself. In the beginning of Section 2, *On the Relativity of Lengths and Times*, Einstein speaks about his two postulates and writes: "These two principles we *define* as follows" (my emphasis; original German: "..., welche beiden Prinzipien wir folgendermaßen definieren."). Needless to say, the formulation of the postulates that follows this introductory sentence is not identical to the formulation Einstein gives of them in other places, even in the same article.

Summing up, there is every reason to believe that there was a considerable amount of constancy and coherence in Einstein's thinking about geometrical concepts and that already in 1905 he was not really proposing to consider determinations by means of rods and clocks as strict definitions of space and time. In fact, it is almost self-evident that spatial and temporal concepts are more fundamental than the existence of rods and clocks: physics does not face any problems in describing imaginary worlds in which there are particles and fields in spatiotemporal configurations, but in which no rods and clocks can exist, not even in principle. Einstein can surely not have believed that space and time loose their meaning as soon as conditions become unfavourable to the existence of rods and clocks. Although the appeal to rods and clocks was well-advised from a strategic and rhetorical point of view, with the purpose of making it clear that classical concepts are not sacrosanct, at a fundamental level space and time should be discussed within the framework of their role in basic physical theory.

Einstein's special theory of relativity served as a beacon for many twentieth-century philosophers of science; but many of them misinterpreted the philosophical implications of the theory (compare [15], for a similar thesis; see also [4], for an early discussion of a related theme). As Howard [15] points out, it was only with the downfall of logical empiricism, and the Quinean criticism of the analytic/synthetic distinction, that philosophy of science caught up with Einstein's thinking about the status of physical concepts. In the philosophy of physics the notion that at the fundamental level it is basic physical theory that is important for the status of space and time, has only rather recently gained substantial ground (cf. [2, 3]). This testifies to the enormous force and persuasiveness of Einstein's rhetorical arguments involving rods and clocks. Nevertheless, the predominance of philosophical misinterpretations remains remarkable, given that an alternative presentation of special relativity has been available for a long time already in the work of Hermann Minkowski.

9.3 Minkowski's Analysis of Space and Time

On 21 September 1908 Minkowski [19] delivered his lecture *Raum und Zeit* (*Space and Time*, [17])[1]. The lecture quickly became exceptionally famous, and the passage from the introductory statement, "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," has become proverbial. Still, I believe that there are a number of aspects of Minkowski's ideas that so far have not received the attention they deserve. I shall here focus on two of them: the ontological status of Minkowski spacetime as a whole, and the meaning of spatiotemporal coordinates, respectively.

With regard to the first issue, I think that the impression created by Minkowski's just-quoted "winged words," namely that he insisted on considering spacetime as an entity existing independently of matter, is mistaken. Rather, examination of the text shows that Minkowski was close to Einstein's sympathies – although it should be borne in mind that Minkowski like Einstein was not explicit on his philosophy of space and time, writing as he was on physics and not on philosophy. Concerning the second issue (the meaning of spatiotemporal coordinates), it is well known that Minkowski did not engage in a discussion of rods and clocks. But exactly how he *did* propose to lay down these coordinates does not appear to have been the subject of serious study in the philosophy of physics literature. This is amazing, for as I shall argue Minkowski's proposed procedure is virtually identical to the procedure Einstein had in mind as fundamental in a future state of physical knowledge. Unlike Einstein, however, Minkowski was not deterred by the fact that he did not actually possess an adequate fundamental theoretical description of macroscopic space-time measuring devices (rods and clocks) and discussed the issue in an abstract way in terms of the form of the physical laws.

Minkowski's leading idea is to start with a theoretical account of elementary physical phenomena in terms of some arbitrary set of variables, then to perform mathematical transformations, and finally to introduce spacetime coordinates as that "system of reference x, y, z, t, space and time, by means of which these phenom-

[1] Many of the ideas can already be found in a lecture Minkowski gave a year earlier (Minkowski, 1915).

ena then present themselves in agreement with definite laws." The "definite laws" Minkowski actually considered were those of Maxwell; but later in his article he assumes that *all* laws of nature, including those yet to be discovered and responsible for the stability of matter, should exhibit the same symmetry properties as the equations of electrodynamics. This procedure, relying as it is on very general symmetry features of the laws of nature, can be regarded as fulfilling Einstein's desiderata, in spite of the absence of complete knowledge of all specific laws and in spite of a lack of insight into how macroscopic measuring devices should be described by fundamental theory – as Minkowski's abstract and elegant mathematical treatment shows, such knowledge is not necessary to complete, albeit in a very abstract form, Einstein's programme.

In the Section 1 of his celebrated paper Minkowski introduces arbitrary coordinates in order to label events: *x, y, z, t*. A set of values of these coordinates represents a "world-point," and the manifold of *all* world-points is the "world." Minkowski makes it immediately clear, however, that he is not thinking of this manifold as an independent entity. First, he stresses that the coordinates are empirically accessible and always occur in union [17, p. 76]: "The objects of our perception invariably include places and times in combination. Nobody has ever noticed a place except at a time, or a time except at a place." Minkowski considers positions and times as attributes of physical, even "observable" ("*wahrnehmbarev*") things. The point is made very explicit a few lines further on, where Minkowski states: "Not to leave a yawning void anywhere, we will imagine that everywhere and everywhen there is something perceptible. To avoid saying "matter" or "electricity" I will use for this something the word "substance." The German original uses the term "*wahrnehmbar*" again: "Um nirgends eine gähnende Leere zu lassen, wollen wir uns vorstellen daß allerorten und zu jeder Zeit etwas Wahrnehmbares vorhanden ist" [19, p. 105]. The use of the term "observable" here may seem strange: Minkowski can hardly be supposed to require that miniscule portions of his substance, whatever its nature, should be accessible to the unaided senses. But we should bear in mind, again, that this is a science paper, not an exercise in philosophy – let alone an application of logical positivist ideas *avant la lettre*. Scientists very often use the term "observable" to denote things or states of affairs that they think are physically existing; things with which it is possible to enter into causal interaction and therefore "indirectly observable." It seems clear that Minkowski's "*wahrnehmbar*" should be interpreted in this vein, and simply should be taken as denoting "physical" or "material".

After his introduction of substantial points, Minkowski focuses attention on the career of one such point, for which he coins the term "worldline." He continues [17, p. 76]: "The whole universe is seen to resolve itself into such worldlines, and I would like to state immediately that in my opinion physical laws might find their most perfect expression as reciprocal relations between these worldlines." ("Die ganze Welt erscheint aufgelöst in solche Weltlinien, und ich möchte sogleich vorwegnehmen, daß meiner Meinung nach die physikalischen Gesetze ihren vollkommensten Ausdruck als Wechselbeziehungen unter diesen Weltlinien finden dürften.") Again, there is no indication here that Minkowski is thinking of his spacetime manifold as something that exists in itself, independently of its material "contents". Rather, his text breathes the atmosphere of Leibnizean, or perhaps rather Machian, relationism. It would not be very consistent, evidently, to capitalize on this point and now, suddenly, to see Minkowski in the role of a philosopher of science. Just as before, we have to read his article as a science text, in which philosophical notions are used in an intuitive and loose way.

Minkowski's view that the laws of physics represent relations between material worldlines is not incidental to his article, however: the idea plays a central role in his subsequent argument. Starting from the laws of physics, which express regularities in the behaviour of material systems, Minkowski starts an analysis from which his spacetime concepts are distilled: the coordinates x, y, z, t are found as those coordinates in terms of which the laws take on their standard forms. Everywhere in this analysis the coordinates refer to physical events, and the properties of spacetime emerge as the invariance properties of the pertinent physical laws. This approach is very different from the usual textbook approach, in which one begins with the symmetries of spacetime and then imposes these same symmetries on the physical laws. Minkowski's method does the opposite thing: one starts with physical systems and the regularities found in their behaviour, then formulates laws and studies the invariance properties of these laws and, finally, one calls the symmetries shared by all physical laws "spacetime" symmetries.

This is how Minkowski himself describes his procedure [17, p. 79]: "From the totality of natural phenomena it is possible, by successively enhanced approximations, to derive more and more exactly a system of reference x, y, z, t, space and time, by means of which these phenomena then present themselves in agreement with definite laws. But when this is done, this system of reference is by no means unequivocally determined by the phenomena. *It is still possible to make any change in the system of reference that is in conformity with the transformations of the group G_c, and leave the expression of the laws of nature unaltered*" (emphasis in the original). The group G_c is the proper Poincaré group, i.e., Lorentz boosts plus translations and rotations. Minkowski had introduced this group a little earlier in his paper, as a purely mathematical guess at a "natural" generalization of the Galilei group. In fact, one of the themes of Minkowski's paper is an ode to the power of pure mathematics to disclose facts about nature – he speculates about the possibility that special relativity could have been discovered by mathematical considerations alone. Possibly this song of praise for mathematics was partly motivated by Minkowski's wish to legitimate his "intrusion," as a mathematician, into a physical research area (cf. [26]). In other places of the paper Minkowski stresses the importance of empirical input, for instance in his famous opening sentence: "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." In the just-described procedure for defining space and time the starting point squarely lies in empirical data. But Minkowski's background as a mathematician remains clearly visible, especially in this prescription for finding inertial spacetime coordinates. Once the data about the relations between the phenomena are in, Minkowski tells us, a mathematical data analysis consisting in rewriting the equations in terms of different systems of independent variables, and successive approximations in order to finally find privileged systems of coordinates, suffices. I think it is not far-fetched to assume that it was exactly his abstract, mathematical, set of mind that made it possible for Minkowski to start with general properties of the physical laws and define spacetime coordinates on that basis. By contrast, Einstein's approach is more typical of the thinking of a working physicist, with a good deal of attention devoted to how spacetime coordinates are actually laid down in practice.

9.4 Intermezzo: Space-Time Coordinates as Physical Properties

According to Minkowski's proposal as explained above, we start with physical phenomena, study their relations, and derive the spacetime description from the resulting equations. In this whole procedure the spacetime coordinates function as attributes of physical systems or physical events, rather than as labels of points in an abstract spacetime manifold. In the context of particle mechanics this suggests treating space and time as physical quantities that have a similar status as mass, charge and similar particle properties. In other words, Minkowski's approach challenges us to think of space-time properties and space-time relations as inhering in physical systems and their relations, and not as referring to properties of an independently existing spacetime manifold (cf. [5, 25]).

The starting point of Minkowski's "successive approximation" method is that physical properties and relations can be quantified in infinitely many ways, depending on choices of scales and units. Given any such choice, the regularities in the dynamical behaviour of the particles assume a particular mathematical form. For example, if we connect particles by means of rods of position-dependent temperature, and express the forces between them in formulas that depend on the number of rods that fit between them (the "distance"), a complicated form of the laws of motion will result, very different from the standard Newtonian or special relativistic one. Conversely, if a standard form of the dynamical laws is given, this imposes restrictions on the way numerical values can be assigned to physical quantities. In particular, the standard position and time values fit in with the standard, inertial, form of the physical laws – with freedom associated with the invariance transformations of the set of equations. The inertial coordinates that emerge in this way are now not defined as markers identifying space-time points, but rather represent a particular way of quantifying space-time properties possessed by the physical systems. In this way of introducing coordinates no need arises to invoke the notion of spacetime as something additional to material systems.

The point deserves further discussion. Particles in classical physics are characterised by a number of intrinsic properties, like mass and electric charge, and by their state of motion. Mass and electromagnetic characteristics of particles are examples of *direct* properties [7]; they inhere in the particles without the need of any intermediary. By contrast, the positions of particles are traditionally viewed as *indirect*. That is, it is traditionally assumed that the geometric relations between the particles (how far apart they are from each other, what their relative orientation is, and so on) derive from the geometric relations between the *spatial points* they *occupy*. Accordingly, the geometric properties of the underlying space are considered as primary and the spatial relations between physical objects as secondary, indirect.

The way I read Minkowski, he proposes to assign space-time properties and space-time relations to particles in the same way as we assign mass values to them. Just as we do not suppose that there is an underlying substantive mass-space in which particles occupy points, we do not need the idea of a container space in which the particles are located. We can consider space-time properties as direct properties, whose introduction is justified by the possibility of formulating physical laws in terms of them. In the same way as mass values play a role in the formulation of regularities in the behaviour of particles, we can use (relative) positions, velocities and accelerations as quantities whose numerical values are needed to formulate laws of motion. As already pointed out, not all assignments of numerical values to the quantity "distance" will lead to the same form of the physical laws. If a preferred form of the laws of motion is specified, only a limited freedom in assigning position values is left.

The traditional objection against the idea that a substantive container space is superfluous is that relationist alternatives to mechanics will not be able to explain inertial effects, like those occurring in Newton's bucket. In those cases both classical and special relativistic mechanics makes a distinction between states of motion that are identical from a relationist point of view as they exhibit the same relative distances, relative velocities, etc., between the particles.

However, this difficulty is not insuperable. In fact, we now know there a purely relational version of classical mechanics exists, that may well be empirically adequate [18] – but it would be anachronistic to associate Minkowski's approach with this possibility. Rather, we should note that even in the context of ordinary classical and special relativistic mechanics it is possible to do justice to inertial effects and to make sense of acceleration without invoking an independent container space-time [5]. For take a way of assigning position and time coordinates such that the dynamics assumes its usual inertial form; if a particle has a non-vanishing second derivative of its position with respect to time *in these coordinates*, it is *accelerated* in the sense needed for the explanation of the bucket experiment and similar examples. Because the mathematical equations are exactly the same as in the usual accounts of mechanics, all the usual results can be reproduced: in particular, a system that is accelerated will evolve differently from a system that is unaccelerated, even if the relative distances and velocities are instantaneously the same. But again, in this approach we can do without an independent space. The distinction between the usual approach and the space-less approach is not in the formalism or in the predictions, but in the interpretation of the formalism. "Absolutely accelerated" in the scheme we are explaining does not mean, "accelerated with respect to absolute space", but rather "with non-vanishing second time derivative of the position according to a coordinate system scale in which the laws of motion have their standard form".

In order to treat time as a direct property as well, on a par with position, it is natural to focus on *events* as the fundamental physical objects with which physical quantities are to be associated. A particle event is assigned three position values (three components of the position) and a time value. As before, transforming these values can subsequently be used to find coordinates that lead to a preferred symmetrical of the dynamical equations.

The important difference between the relativistic and the Newtonian setting is that in relativity the symmetry transformations that leave the symmetric form of the equations invariant mix space and time quantities. What is invariant in special relativity is the *space-time interval* between two events, and relativistic particle dynamics can be formulated completely in terms of this four-dimensional distance. For example, the free motion of a particle is such that it makes the four-distance between the events in its existence a maximum. In the relativistic context the direct property view of space and time therefore assumes the following form. We start with particle events as our physical "objects"; they are assigned coordinates, i.e., four numbers x, y, z, t. These coordinates are then transformed in such a way that the relativistic equations of motions hold in their inertial form. This still leaves a lot of freedom, because the dynamics is invariant under the transformations of the Poincaré group. To make a connection with the conventional approach, one can think of these various assignments of position and time values as the result of applying standard measuring procedures from different inertial frames of reference.

The transition to relativity thus makes it possible to satisfy a traditional relationist desideratum, namely to make all quantities *relative*, in the sense of pertaining to relations between physical objects. However, the physical "objects" we are speaking about now, in the relativistic context, are particle events, and not the particles-at-the-same-time of Leibnizean relationism. This change of perspective results in a reconciliation between the relative character of the basic quantities, and the absoluteness of being accelerated. This is because Lorentz transformations not only leave the four-dimensional interval ds invariant, but also preserve the distinction between being accelerated and being in an inertial state of motion.

In the case of fields, instead of particles, there is the complication that fields seem to require the prior existence of the continuum of coordinate values for their very definition: fields are standardly defined via the assignment of field values to space-time points. Some authors (cf. [7]; ch. 8) take this to imply that field theories need an independently existing space-time manifold for their very possibility of existence: an assignment of properties to space-time points obviously requires the assumption that space-time points exist. However, this argument is inconclusive. Of course, *if* properties are indeed assigned to *space-time points*, the assumption is that there *are* such space-time points; this is tautological. The real question is whether the assignment of field values to x, y, z and t, without assuming anything beforehand about what these coordinates refer to, necessarily implies that they refer to independent space-time points. It seems obvious that the answer to this latter question is in the negative. If quantities are represented as functions of certain coordinates, it clearly does not follow that these coordinates refer to something real, existing independently of the things that are being coordinatized. The instance of colours and their mutual relations furnishes an example. Different colours and their shades can be represented in various ways; one way is as points on a three-dimensional colour solid. But the proposal to regard this "colour space" as something substantive, needed to ground the concept of colour, would be strange to say the very least. Of course, it is exactly the other way around: colours have certain relations among themselves, and their comparison makes it expedient to introduce the notion of a colour space. Coordinates are introduced to mathematically handle this ordering scheme, and to represent the colour relations. Exactly the same idea, transplanted to the context of space and time, lies at the basis of Minkowski's proposal: we start with physical systems and events, systematize their relations, and introduce space time coordinates to mathematically represent the result of this.

9.5 Simultaneity, Symmetries and Conventionality

We have seen that Einstein introduced his rule for establishing simultaneity with the term *definition*; this has been seized upon by many later philosophers of physics, to invoke Einstein's authority for the thesis that simultaneity in special relativity is conventional. Reichenbach [21] gave a systematic elaboration and explanation of this idea that has been very influential. The core of the conventionality doctrine is that local clock indications are objective because they consist in coincidences of material objects, like the coincidence of a pointer with a mark on a dial. These are things we can immediately observe and that do not depend on conventions (except in the trivial sense that we think up words to describe them, choose units in order to number the marks on the dial, and decide to fix our attention on these things in the first place). Whether or not the hands of a clock touch a certain mark on the clock's dial is something given by nature, and not determined by us. By contrast, simultaneity cannot be directly perceived: we need some rule to tell us how to establish simultaneity on the basis of observable facts, and it is only the stipulation of this rule that gives content to the concept. Such stipulations can be made in different ways, which gives rise to different but equivalent descriptions. The differences between these descriptions obviously are differences in judgments of simultaneity; the equivalence consists in the fact that the local states of affairs, the coincidences, remain the same in all of them. This epitomizes the philosophical, empiricist interpretation of Einstein's use of "definitions": local coincidences are objective facts, all the rest is a matter of choice. Of course, some choices may be simpler than others in the rules they use, or lead to simpler laws. But this can only yield a *pragmatic* argument for preferring one definition over another. Such pragmatic arguments relate to our interests and preferences, but not to truth. Thus, Reichenbach admits that the definition of simultaneity that makes the speed of light the same in all directions and leads to the standard form of Maxwell's equations is simpler for us, easier to remember and more readily applicable than alternatives. But according to Reichenbach that does not at all mean that these alternatives (with a value of ε unequal to $^1/_2$) have a lesser claim to being true. As long as the local facts remain the same, all theoretical schemes that accommodate them are equally true or false. Reichenbach does note that it is an objective fact that a choice of simultaneity that leads to isotropy of the speed of light is *possible*; but he emphasizes that still, it is our conventional choice to make use of this circumstance and set $\varepsilon = {}^1/_2$.

The conventionalist thesis thus depends on the notion that only local states of affairs are physically objective. For the logical empiricists this accorded with central tenets in their philosophy of science, because such local facts are paradigmatic of what is directly accessible to the senses. They were inspired by special relativity and Einstein's pronouncements; and indeed, as we have seen, Einstein himself made statements that appear to go into the same direction – in his explanation of what we have to understand by "time," Einstein starts with the uncontroversial local times indicated by clocks, and only subsequently "defines" a global time via his recipe for synchronizing clocks. But the appeal to Einstein's authority on this point is misdirected. As we have stressed before, there is every reason to suppose that Einstein did not intend his remarks as a consequence of a systematic empiricist philosophy of science, but rather as a stratagem employed to convince his readers that the temporal structure of the universe might be very different from that assumed in classical mechanics. Einstein wanted to make clear that on close inspection it turns out that there is no empirical support for the classical notion of absolute simultaneity. His target was not the uniqueness of the synchronization rule, but rather the tenability of the *classical* conception of time. Indeed, he never even considered using other definitions than his standard one for establishing simultaneity; and what is more, he directly linked this standard definition to the resulting form of the physical laws. In the beginning of Section 2 of his 1905 paper, when after the preliminary remarks of Section 1 he starts addressing the physical content of relativity theory, he again formulates his two basic principles, the relativity principle and the light principle. The latter is now formulated as: "Any ray of light moves in the 'stationary' system of co-ordinates with the determined velocity c, whether the ray be emitted by a stationary or by a moving body." Einstein immediately adds that in the definition of velocity, time must be taken in the sense of the definition of his Section 1. In other words, the way time has been defined in Section 1 of the 1905 paper is precisely such that the light principle receives its usual and natural form, with equal speed of light in all directions. Moreover, even *before* discussing the "meaning of time" in Section 1, Einstein had formulated the two relativity postulates in his Introduction, and had already worded his light postulate as the principle saying that light is always propagated in empty space with a definite velocity c, independent of the state of motion of the emitting body. This is about light in empty space (apart from the emitting body and the light itself), without any clocks, rods, and without signals going to and fro. The subsequent discussion in Section 1 has the purpose of making this postulate understandable and consistent with the other postulate. In other words, in Einstein's construction of the theory of special relativity the light postulate comes before the definition of time. Seen this way, Einstein's approach is similar in spirit to Minkowski's theory-centred one.

In contradistinction to Einstein, Minkowski explicitly states that he starts with physical theory in his construction of the spacetime manifold. The coordinates x, y, z, t are determined by Minkowski as the coordinates in terms of which the theory assumes a preferred form (namely the standard one, which among other things makes the speed of light isotropic). Of course, this physical theory is not given *a priori* to us. In his schematic description of his proposed procedure Minkowski explains that the theory should be distilled from regularities in the observed phenomena. That means that also Minkowski begins with local observations; but from the very start he takes into consideration what physical *relations* exist between these local phenomena, how they compare, and what regularities there are in the global pattern of local phenomena. Put differently, from the start Minkowski's approach includes *global* aspects of the situation: the spacetime coordinates that he constructs are sensitive to global properties of the pattern of events. This is especially true for the notion of simultaneity that follows from Minkowski's construction. Minkowski's simultaneity relation reflects the isotropy and homogeneity of spacetime, in the sense of the symmetry properties of the physical laws. As admitted even by Reichenbach, this isotropy and homogeneity is an objective physical *fact*: it proves possible, as an empirical result, to *find* a consistent description in which the laws display identical properties at all points in space and time, and in all directions (this is the description with $\varepsilon = {}^1/_2$). This existence claim is obviously not *a priori* true – in fact it is false in most general relativistic spacetimes. That it is true in special relativity tells us something objective and important about the nature of the continuum of events: it is highly symmetric. Standard simultaneity reflects this global symmetry by not only making the velocity of light a universal constant, but by making *all* fundamental physical processes that propagate in time position and direction independent.

So within the framework of Minkowski's approach standard simultaneity is not conventional but represents an objective physical fact. To some extent this argument was already anticipated by Reichenbach – and it seems that Reichenbach did not feel completely secure about whether Einstein was really behind the conventionality thesis. Indeed, apparently in order to dissociate himself from possible Einsteinian reservations on exactly this point, Reichenbach [21, p. 124] wrote: "Einstein immediately applied his solution of the problem of simultaneity to theoretical physics and for this reason the epistemological character of his discovery has never been clearly distinguished from the physical results. Therefore, we shall not follow the road taken by Einstein, which is closely connected with the principle of the constancy of the velocity of light, but begin with the epistemological problem."

It is exactly at this point that the misinterpretation of the philosophical message of special relativity begins to take concrete shape. Indeed, a couple of pages later, Reichenbach [21, p. 127] introduces his famous ε-formula, and after duly mentioning that only the value $\varepsilon = {}^1/_2$ was considered by Einstein, he continues: "*This definition* (i.e., with $\varepsilon = {}^1/_2$ – my addition) *is essential for the special theory of relativity*, but it is not epistemologically necessary. Einstein's definition, too, is just one possible definition" (emphasis added). So Reichenbach recognizes the essential role played by the standard notion of time in special relativity; but he denies that this special role has anything to do with objective facts of nature. Indeed, he goes on to state that Einstein's preference for $\varepsilon = {}^1/_2$ is solely based on the fact that this choice leads to simpler relations, and says: "It is clear that we are dealing here merely with descriptive simplicity, the nature of which will be explained in §27." Rather surprisingly, in §27 we find Reichenbach again emphasizing that a special definition of simultaneity is possible, precisely the one with $\varepsilon = {}^1/_2$, and that this special definition possesses important advantages because it makes the simultaneity relation symmetric and transitive. As Reichenbach stresses in the same passage, the existence of such a special definition is by no means self-evident but requires specific physical facts. That these requirements are indeed fulfilled in special relativity is what justifies Einstein's definition. But, Reichenbach continues [21, p. 168]: "This should not mislead us into believing that this definition is 'more true' because of its simplicity. Again we are concerned with nothing but descriptive simplicity." This is disappointing: we were promised an explanation of why $\varepsilon = {}^1/_2$ has only pragmatic virtues, which do not relate to truth; but here the earlier claim to that effect is just repeated. However, within the general context of Reichenbach's philosophy of science the motivation of the judgment is clear enough: for Reichenbach only local coincidences are objective building blocks of scientific theory, whereas the way we describe their correlations is conventional. Within this conceptual framework any advantage one global description possesses over another can necessarily only relate to pragmatic factors like simplicity, elegance, beauty, and so on.

The thesis defended by Reichenbach, and by many others in his wake, thus boils down to the following. There *are* global symmetries in Minkowski spacetime, and it *is* true that these are best represented by standard simultaneity (which is, of course, relative to a state of inertial motion). For example, the velocity of light will only conform to the symmetry and have the same value in all directions if we accept this standard simultaneity. But still, it is a conventional choice to *exploit* this possibility. But isn't this like saying that it is true that there are macroscopic objects in our world, that it is also true that this state of affairs is fittingly represented by a language that refers to these objects, but that it is still a matter of conventional choice to actually opt for such a language? In other words, is this conventionality not just the trivial conventionality that follows from the fact that *we* decide to use a language, that *we* coin words, and so on? Put differently again, is it not true that this kind of conventionality does not follow from the non-existence of relevant physical facts pertaining to simultaneity, but rather from a strategy that can be applied across the board and makes, if consistently employed, *every* physical concept conventional? As we have seen, in addition to local facts there exist *global* ones, as is admitted by all parties concerned. In special relativity there exist objective global spacetime symmetries, and it seems a highhanded measure to dismiss them as unimportant for the notion of simultaneity. If these global facts are taken into account, there is no reason to deem standard simultaneity in special relativity conventional (the situation is different in general relativity, or in accelerated frames of reference – indeed, in those contexts, in which there are no global symmetries, the case against the objectivity of global simultaneity becomes much stronger – see [6]).

Concluding, I think that Minkowski hit the nail on its head when he analysed space and time as implicitly defined by physical theory, and that by doing so he made explicit what was implicit in Einstein's original approach. If this is correct, it follows that neither from Einstein's nor from Minkowski's work support can be derived for the existence of the "epistemological revolution" that the logical empiricists perceived in relativity theory. In particular, the notorious conventionality thesis that was so ardently defended by many philosophical commentators on Einstein's revolution appears as a consequence of philosophical prejudices rather than as a part of relativity theory.

References

1. Bridgman, P.W.: 'Einstein's Theories and the Operational Point of View', pp. 333–355 in Schilpp (1949)
2. Brown, H.R.: Physical Relativity. Clarendon, Oxford (2005)
3. Dieks, D.: The "reality" of the Lorentz contraction. Zeitschrift für allgemeine Wissenschaftstheorie **15**, 33–45 (1984)
4. Dieks, D.: Gravitation as a universal force. Synthese **73**, 381–397 (1987)
5. Dieks, D.: Space and time in particle and field physics. Stud. Hist. Phil. Mod. Phys. **32**, 217–242 (2001)
6. Dieks, D.: Space, time and coordinates in a rotating world. In: Rizzi, G., Ruggiero, M.L. (eds.) Relativity in Rotating Frames, pp. 29–42. Kluwer, Dordrecht (2004)
7. Earman, J.: World Enough and Space Time. MIT Press, Cambridge, MA (1989)
8. Einstein, A.: Zur Elektrodynamik bewegter Körper. Annalen der Physik **17**, 891–921 (1905)
9. Einstein, A.: Über das Relativitätsprinzip und die aus demselben gezogen Folgerungen. Jahrbuch der Radioaktivität und Elektronik **4**, 411–462 (1907)
10. Einstein, A.: Le Principe de Relativité et ses Conséquences dans la Physique Moderne. Archives des sciences physiques et naturelles **29**, 5–28/125–144 (1910)
11. Einstein, A.: Die Relativitätstheorie. Naturforschende Gesellschaft in Zürich. Vierteljahrsschrift **56**, 1–14 (1911)
12. Einstein, A.: Geometrie und Erfahrung. Julius Springer, Berlin (1921)
13. Einstein, A.: Nichteuklidische Geometrie und Physik. Die neue Rundschau **1**, 16–20 (1925)
14. Einstein, A.: Ideas and Opinions. Crown Publishers, New York (1954)
15. Howard, D.: 'Einstein and the Development of Twentieth-Century Philosophy of Science', to appear in *The Cambridge Companion to Einstein* (2007)
16. Klein, M.J., Kox, A.J., Renn, J., Schulmann, R. (eds.): The Collected Papers of Albert Einstein, vol. 3. Princeton University Press, Princeton, NJ (1993)
17. Lorentz, H.A., Einstein, A., Minkowski, H., Weyl, H.: The Principle of Relativity. Methuen, London (1923) (First republished as a Dover edition in 1952, Dover, New York)
18. Lynden-Bell, D.: A relative Newtonian mechanics. In: Barbour, J., Pfister, H. (eds.) Mach's Principle from Newton's Bucket to Quantum Gravity, pp. 172–178. Birkhäuser, Basel (1995)
19. Minkowski, H.: Raum und Zeit. Physikalische Zeitschrift **10**, 104–111 (1909)
20. Reichenbach, H.: 'The Philosophical Significance of the Theory of Relativity', in Schilpp (1949), pp. 287–313 (1949)
21. Reichenbach, H.: The Philosophy of Space and Time. Dover, New York (1957) (Original German version: Philosophie der Raum-Zeit-Lehre. Walter de Gruyter, Berlin (1928))
22. Schlick, M.: Space and Time in Contemporary Physics. Oxford University Press, Oxford (1920) (Original German version: Raum und Zeit in der gegenwärtigen Physik. Julius Springer, Berlin (1917))
23. Schilpp, P.A. (ed.): Albert Einstein: Philosopher-Scientist. Open Court, La Salle (1949)
24. Stachel, J., Cassidy, D.C., Renn, J., Schulmann, R. (eds.): The Collected Papers of Albert Einstein, vol. 2. Princeton University Press, Princeton, NJ (1989)
25. Teller, P.: Space-time as a physical quantity. In: Achinstein, P., Kagon, R. (eds.) Kelvin's Baltimore Lectures and Modern Theoretical Physics. MIT Press, Cambridge, MA (1987)
26. Walter, S.: Minkowski, Mathematicians, and the Mathematical Theory of Relativity, pp. 45–86. In: Goenner, H., Renn, J., Ritter, J., Sauer, T. (eds.) The Expanding Worlds of General Relativity. Birkhäuser, Boston/Basel (1999)

Chapter 10
Hermann Minkowski: From Geometry of Numbers to Physical Geometry

Yvon Gauthier

Abstract For the historian or philosopher of science, Hermann Minkowski is known for the formulation of Special Relativity in terms of four-dimensional space-time. The original text is the famous 1908 *"Raum und Zeit"*, but it is rarely mentioned that Minkowski is the author of a geometry of numbers *"Geometrie der Zahlen"*, a most important work in number theory. In his arithmetic geometry, Minkowski introduces the notion of numerical grids or lattices (*Zahlengitter*) that are meant as a geometrical representation of arithmetical relations, that is isolated points and intersection points used to define the approximation of a real number by rational numbers.

I want to show that the concept of a numerical grid is the origin of Minkowski's diagrams in the physical geometry of Special Relativity. Minkowskian spacetime is isomorphic to a universal numerical grid with no ontological import. This might also be relevant for the new physics of spacetime (with strings or loops) which puts into question the modern concepts of space and time in their relativistic or quantum-mechanical settings.

Keywords Number theory · Geometry · Physics · Physical axiomatics

10.1 Introduction

In the present paper, I am interested in the connection of Minkowski's spacetime formulation of Special Relativity with his geometry of numbers or *Geometrie der Zahlen* which Minkowski had developed prior to his famous 1908 paper *"Raum und Zeit"*. There is an inner mathematical connection between the two enterprises and I want to argue that spacetime diagrams are an illustration in physical geometry of a central scheme in the geometry of numbers which was the main endeavour of Minkowski's mathematical career. Minkowski's work in physics belongs to mathematical physics and his statement at the end of the *"Raum und Zeit"* paper can be counted as the philosophical motto of the mathematical physicist:

Y. Gauthier (✉)
Université de Montréal, Montréal, QC, Canada, H3C 3J7
e-mail: yvon.gauthier@umontreal.ca

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_10,

> With the elaboration of its mathematical consequences, there will be plenty of hints for the experimental confirmation of the postulate (of the absolute world), so that anyone who feels uncomfortable with the loss of traditional pictures (*Anschauungen*) will find himself compensated with the idea of a preestablished harmony between physics and pure mathematics. [5] (my translation)

Such a declaration of principle echoes Minkowski's assessment of Dirichlet's achievements in mathematical physics – see his 1905 address *"Peter Gustav Lejeune Dirichlet und seine Bedeutung für die heutige Mathematik"* (in [5] II:449–461). Minkowski says that the two directions of number theory and mathematical physics, though they seem to diverge, are harmoniously integrated in Dirichlet's work by the use of the integral calculus. Here Minkowski mentions Dirichlet's results on the convergence of Fourier series and the Dirichlet principle on the minima of the potential function. In praising Dirichlet for having introduced discontinuous factors in the multiple integrals of the potential function, Minkowski evokes Leibniz's idea of a grand scheme for a perfectly harmonious world. This mathematician's dream was encapsulated in Minkowski's postulate of the absolute world and I want to explore now its mathematical motivation.

Hermann Weyl credited Minkowski for having recognized that:

> The fundamental equations for moving bodies are determined by the principle of relativity if Maxwell's theory for matter at rest is taken for granted. [7]

He also said, referring to Minkowski's 1907 paper *"Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körper"* (in [5] II:352–404) or "The Fundamental Equations for the Electromagnetic Processes in Moving Bodies", that:

> The adequate mathematical formulation of Einstein's discovery was first given by Minkowski: to him we are indebted for the idea of four-dimensional world-geometry. [7]

Minkowski had distinguished in the aforementioned paper the theorem of relativity from the principle of relativity; the first is purely mathematical in terms of the covariance of Lorentz transformations and the second, the principle of relativity, allows, in Minkowski's words, for the derivation of the laws of mechanics solely from the principle of the conservation of energy. The language here is still of space-time vectors and does not anticipate on the 1908 paper on *"Raum und Zeit"* where the vocabulary of worldpoints and worldlines is canonized and serves as the basic ingredient for a graphic representation, as Minkowski says, of the group of spatiotemporal transformations. Minkowski then suggests that the terminology for the postulate of relativity is rather dull – *"matt"* in German – when one wants to stress the invariance properties of the group of the transformations and he comes up with his "postulate of the absolute world" (*Postulat der absoluten Welt*). Let me point out that at one time Einstein himself had wanted to rename relativity theory as invariant theory.

My contention is that Minkowski's idiom has essentially a mathematical meaning as a representational means for the spatiotemporal structure of the physical world. And I assume that Weyl did not interpret the Minkowskian mathematical picture otherwise. I shall give some reasons to substantiate that claim in the following.

10.2 Geometry of Numbers

Minkowski's work on physical problems (hydrodynamics, capillarity, etc., and on relativity theory) is marginal compared to his endeavour in number theory, geometry and especially what Minkowski called the geometry of numbers. It is in the geometrical representation of number-theoretic relations that Minkowski introduces the notion of *Zahlengitter* or number grids. Minkowski defines there a three-dimensional number grid as a geometrical representation of three integers in rectangular coordinates – see *"Über Geometrie der Zahlen"* in [5] I:264–265. These three integers correspond to discrete points in space; and these points in turn represent bodies (*Körper*) and the main question is the content of the surface (*Flächeninhalt*) on which those bodies float or are immersed in, so to speak (Fig. 10.1). To illustrate this train of thought, let us look at a diagram drawn by Minkowski to define the approximation of a real quantity (number) by rational numbers:

The three grid points H, J and K on the straight line $Y = 1$ stand in the following relation: $HJ = JK = OA$ for O the null point at the center of the grid and A a point corresponding to K on the straight line G parallel to $Y = 1$. The inscribed parallelogram does not contain any other grid point besides O in its interior. Minkowski uses such diagrams to show that two relatively prime numbers x and y can be represented by linear forms $\zeta = \alpha x + \beta y$, $\eta = \gamma x + \delta y$ for the straight lines ζ and η with arbitrary coefficients $\alpha, \beta, \gamma, \delta$ and a determinant $\alpha\delta - \beta\gamma = 1$ so that the norm or the length between x and y (not both zero) is

$$|\zeta\eta| \leq 1/2.$$

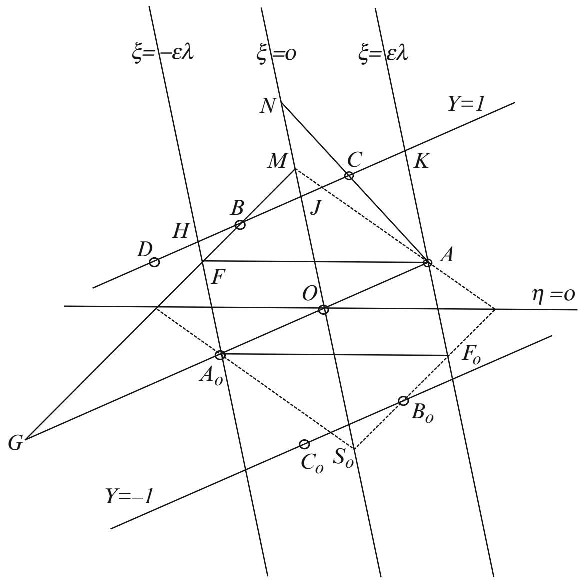


Fig. 10.1 Number grid (From Minkowski [5] I:327)

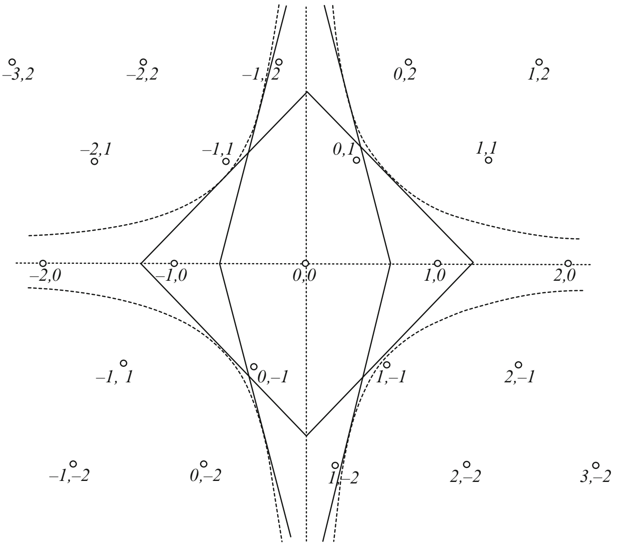
Minkowski's diagrams and calculations are quite involved as is shown by the following diagram intended to illustrate the "very intuitive link between all the possible solutions of the inequality $|\zeta\eta| \leq 1/2$ in whole numbers without common divisors" in Minkowski's words (Fig. 10.2):

Here Minkowski uses continued fractions to obtain the desired inequality

$$-1/2 < \zeta\eta < 1/2$$

for what he calls diagonal chains of continued fractions (*Diagonalkettenbrüche*).

This geometry of numbers has an arithmetical core, while geometry has an intuitive appeal. The main idea is to inscribe triangles or parallelograms in rectangular Cartesian coordinates in order to represent geometrically the reticular system or grid (*Gitter*) of all grid points (*Gitterpunkte*) of positive quadratic forms with integer coefficients. A number grid or mesh or point lattice (as it is now called when points



(1) $$f(z) = c_m z^m + \ldots + c_0 + \frac{c_1}{z} + \frac{c_2}{z^2} + \ldots = F_0(z) - \cfrac{1}{F_1(z) - \cfrac{1}{F_2(z) - \cdots}};$$

(2) $$\big(P(z) - f(z)\, Q(z)\big)\, Q(z).$$

(3) $$\frac{x}{y} - a < \frac{1}{2\, y^2}, \qquad (4) \quad a = g_0 - \cfrac{1}{g_1 - \cfrac{1}{g_2 - \cdots}}$$

(5) $$\xi = \alpha x + \beta y,\ \eta = \gamma x + \delta y,\ \alpha\delta - \beta\gamma = 1$$

(6) $$-\tfrac{1}{2} < \xi\eta < \tfrac{1}{2}.$$

Fig. 10.2 *"Diagonalketten"* (diagonal chains) (From Minkowski [5] II:45) have integer coefficients) is most important for the representation of the volume of a body and its fundamental arithmetical property is the generalization of the length of a straight line into the principle that in a triangle the sum of the lengths of two sides is never smaller than the length of the third side. As a special case, one easily points to the Pythagorean Theorem for right triangles:

$$c^2 = a^2 + b^2$$

which is at the foundation of the differential form

$$ds^2 = dx_1^2 + dx_2^2$$

in Euclidean coordinates. Weyl pointed out that the differential form ds^2 is not only the simplest, but also the canon for the classification of possible geometries since the positive quadratic form generates all linear transformations of the variables involved as a unique mathematical structure. Of course, the ds^2 is the fundamental (quadratic) form

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$$

for the invariant metric element in Special Relativity. Let me remark that if one lifts the quadratic restriction, one gets Finsler spaces which are akin to Riemannian spaces. I do not need to mention that Minkowski devoted much of his work to the theory of quadratic forms, i.e., homogeneous polynomials of the second degree, which were the main object of study in number theory from Gauss to Kronecker. Rather than elaborate on this, I can only mention that number grids are clearly connected to what we now call Minkowski diagrams.

10.3 Spacetime Diagrams

In his mathematical diagrams, Minkowski depicts grid points and intersection points for the approximation of a real number (or quantity) by rational numbers or diagonal chains of continued fractions that illustrate the fact that in a triangle the sum of the lengths of two sides cannot be smaller than the length of the third one. Here entire rational functions are used as denominators of continued fractions. What we have is an intuitive representation of solutions of a real inequality in terms of integers without common divisors.

Those diagrams do not differ essentially from the ones which are to be found in the paper *"Raum und Zeit"*. What is represented here as a grid point is the notion of an arbitrary point like charge or electron or potential field (with vector potential and scalar potential components) in a light cone. The world postulate or the postulate of the absolute world is nothing more than the totality of those grid points along grid lines in a gridded universe without any ontological import.

To obtain a physical picture for Minkowski's diagrams, one has to make the assumption of a correspondence between the energy vector *of motion* and the energy vector *in motion*, as Minkowski clearly said:

> *Der Kraftvektor der Bewegung ist gleich dem bewegenden Kraftvektor.* [5] II:441

What this means is that motion can only be represented by the picture of a moving vector on a continuous line, a worldline as a moving world point; diagrams can then be drawn to picture motion in a physical geometry as grids were used to cover the content of a surface (*Flächeninhalt*) in a geometry of numbers (arithmetical geometry). The parallel between the two tasks, to cover a surface with numerical grids and to fill up a two-dimensional space with diagrams, strongly suggests that there is a continuous path in Minkowski's mathematical *Weltanschauung* or theoretical construction of the world, to use Weyl's terminology. Let us examine in detail the following diagram (Fig. 10.3). The diagram depicts an electron e in motion along its worldline through worldpoints P, $P1$ and Q. Minkowski wants to describe here a moving pointlike charge in the "absolute world", that is in a four-dimensional continuum with three space coordinates x, y and z and a time coordinate t. The vector $P1Q$ has a norm or length r, whereas the vector PQ has a length e/r since P lies on the tangent orthogonal to $P1Q$ which cuts through the worldline of e. All this combined with c, the speed of light in the light cone, defines the potential field of the pointlike electron e at point $P1$. Set against the negative curvature hyperbola of the timelike world line, the diagram should provide the appropriate scene for the equations of the ponderomotive force in an electromagnetic field.

Let us remark that this diagram is meant to represent a four-dimensional world – there are continuous curved lines – while the diagrams in the geometry of numbers were meant to represent a three-dimensional world of spatial bodies (*Körper*) –

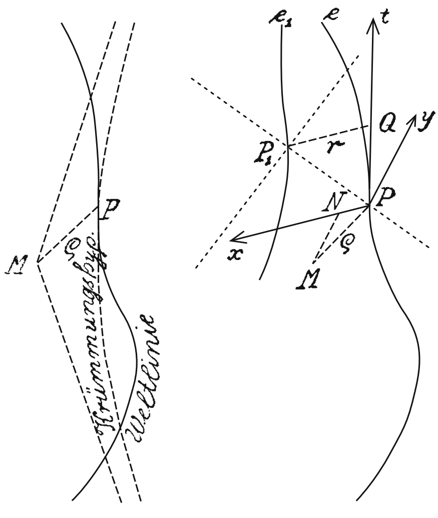


Fig. 10.3 Spacetime diagram (From Minkowski [5] II:442)

there were only continuous straight lines. The four-dimensional world is of course expressed in the quadratic form

$$dt^2 = -dx^2 - dy^2 - dz^2 - ds^2$$

or equivalently for the invariant metric element

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$$

which is the central formulation of Minkowski's version of Special Relativity.

10.4 Physical Axiomatics

Is this Minkowskian representation a physical or a mathematical one? In view of Minkowski's pronouncements on mathematical physics, one would be tempted to advocate a purely mathematical treatment. This might be the reason why Einstein was reluctant at first to adopt Minkowski's formulation. I would rather put Minkowski's achievements in the Hilbertian tradition of the axiomatization of physics. We know that Hilbert was a close friend of Minkowski's and he praised him on many occasions.

One is reminded that Hilbert had put the problem of "The mathematical treatment of the axioms of physics" as the number 6 item on his famous 1,900 list of mathematical problems. Hilbert names probability theory and mechanics as the two major candidates for axiomatization. The central problem in physical theories is the consistency problem in Hilbert's view because a fundamental physical theory proceeds like geometry from general axioms to more specific ones and the extension from the first principles to the secondary ones must preserve consistency. Consistency is not a matter of feeling or experimentation, but of logic, Hilbert insists.

The problem area under discussion is of no particular interest for our purposes, nor are Hilbert's contributions to relativity theory (in [4] III:257–289) since they are mathematical elaborations and only partly foundationaly illuminating – Hilbert had also worked on the foundations of the kinetic theory of gases and other occasional physical subjects. The work on (general) relativity theory in particular seems to have been inspired by the groundbreaking inquiries of Weyl, more than by Einstein's original work. Of greater interest to us is the paper written in collaboration with von Neumann and Nordheim on the foundations of quantum mechanics [3].

In that paper we find the clear exhortation to make explicit the concept of probability in order to extract the mathematical content from its mystical (philosophical) gangue. But the main themes are, in my view, associated with the notions of "analytical apparatus" (*analytischer Apparat*) and "conditions of reality" (*Realitätsbedingungen*). Which comes first, the analytical apparatus or conditions of reality, is a matter of foundational outlook and we shall see how Hilbert conceived of a so-called "physical axiomatics".

Probabilities and their relationships constitute the material we start from. The physical requirements a probability theory of physical phenomena has to fulfil represent the basis on which a "simple" analytical apparatus is defined; then follows a physical interpretation of the analytical structure, and if the basis is fully determined, the analytical structure should be canonical. This is the axiomatic formulation already present in the Hilbertian foundations of geometry and the general argument leaves no doubt as to the permanence of the axiomatic ideal in Hilbert's work on the foundations of physics. What Hilbert seems to strive for is the conception of a categorical mathematical theory with a multiplicity of models; however, not all models would be isomorphic. Non-standard models point rather to a complete first-order theory that generates a variety of interpretations. But the mathematical structure is generally not first-order. The dilemma of a physical axiomatics or of a "physical logic" opens up numerous avenues of research.

The analytical apparatus or the mathematical formalism is first conjectured and then tested through an interpretation in order to check its adequacy. The two components, analytical apparatus and its physical interpretation, must be sharply distinguished and that separation has the effect that the formalism is stable throughout the variations of its (physical) interpretations where some degree of freedom and arbitrariness cannot be eliminated. However, this is the price to pay for the axiomatization and vague concepts like probability will finally loose their fuzzy character. The conditions of reality for probability will prove to be intrinsically linked with the calculus of Hermitian operators and Hilbert's early theory of integral equations. Thus the fact that a probability measure is real positive depends on the finiteness of the sum

$$a_1x_1 + a_2x_2$$

for a linear function. Hilbert's result, which is a building-block of the Hilbert space formalism, was inspired by a similar result of Kronecker on linear forms. Kronecker's influence on Hilbert has also a conservative extension in the foundations of quantum mechanics and not only in the foundations of mathematics [1].

Hilbert's ideas of the foundations of Quantum Mechanics have been made to work by von Neumann in the Hilbert space formulation, which is the standard formulation of Quantum Mechanics. From my point of view, Minkowski's diagrams belong to the models of a canonical analytical apparatus and I want to come back to Minkowski's other works in mathematical physics. I see Minkowski's endeavour as part of the Hilbertian program for the axiomatization of physics. Hilbert had mentioned as candidates for axiomatization limit processes and laws of motion for solid bodies in continua: it is precisely that kind of problem that Minkowski had already tackled in a 1888 paper on "The Motion of a Solid Body in a Liquid" ("*Über die Bewegung eines festen Körpers in einer Flüssigkeit*") in the line of Kirchoff's work. His subsequent work on capillarity is a sequel to that first work and it is in 1907 paper mentioned above on "The fundamental equations for electromagnetic processes in moving bodies" that he states his relativity theorem (*Theorem der Relativität*) for the pure mathematical fact of the covariance of Lorentz transformations and then proceeds to formulate his relativity principle in a canonical (axiomatic) fashion, for which everything follows from the sole principle of the conservation of energy. In that paper, Minkowski declares that

> It is not a serious difficulty for the mathematician accustomed to *n*-dimensional manifolds and non-Euclidean geometry to adapt the concept of time (as a fourth dimension) to the concrete Lorentz transformations. [5] II:366 (my translation).

This is in accordance with Einstein's account of Special Relativity, Minkowski concludes. Minkowski's efforts in the axiomatization of physics are continued in the paper on "The derivation of the fundamental equations for electromagnetic processes in moving bodies from the point of view of the theory of the electron" [5] II:405–430). But that text had to be rewritten by Max Born on Hilbert's invitation after Minkowski's death. Born says that he had to work with a hundred-page manuscript full of formulas but with no helpful advice; needless to say, it is not a conclusive treatment of the theory of electrons.

10.5 Conclusion

What philosophical conclusions can be drawn from my analysis? For the ontology of space-time, a grid universe could be empty or devoid of any substance, or *Substanz* if we want to use Minkowski's word for matter. Substantial points or worldpoints constitute worldlines which embrace the whole world, Minkowski insisted (in [5] II:434) and the universal validity of the world postulate points to the pre-established harmony of pure mathematics and physics (in [5] II:444). Since the postulate of the absolute world can be given a purely mathematical signification in virtue of its arithmetico-geometrical foundations, I would grant it, following Weyl, only a transcendental status, that is the status of an a priori structure in the theoretical construction of the world, as Weyl put it in his *Philosophy of Mathematics and Natural Science* (in [6] p. 235). Minkowski diagrams belong to the analytical apparatus – *der analytische Apparat* as Hilbert had termed it in his work on the foundations of Quantum Mechanics [1]. The diagrams do not belong to a model of the physical world. From that standpoint, a realist interpretation of the Minkowskian world view is thereby excluded on Minkowski's own terms. Although Minkowski states that his intuitions of space and time rest on the solid ground of experimental physics, their validity must be sought in the mathematical justification of the intuitive content. Behind or below the physical geometry, *"physikalische Geometrie"*, as Helmholtz called it, lies a geometry of numbers, the heart of which is arithmetic or number theory or simply number. Minkowski's four-dimensional manifold is an instance of the concept of *n*-dimensional manifold and the mathematician has no problem in dealing – or toying, one might say – with that notion.

The concept of an *n*-dimensional manifold or space (*Mannigfaltigkeit*) is a topological one, that is, it is to be treated with continuous (real and complex) functions. As a metric space, it is endowed with the metric invariant ds^2 which Riemann refers to Gauss' theory of curved surfaces – for Weyl's treatment of Riemannian geometry, see my paper [2].

The fundamental quantity for any surface is the element of arc length

$$ds^2 = dx^2 + dy^2 + dz^2.$$

Gauss had given a general form to this notion of distance by introducing parameters u and v to represent the coordinates $(x,\ y,\ z)$ of any point on the surface and had obtained

$$ds^2 = e(u,v)dv^2 + 2f(u,v)dudv + g(u,v)dv^2$$

(where $e,\ f,\ g$ are functions of the parameters u and v). The general formula for the ds^2 is

$$ds^2 = g_{mn}dx_m dx_n$$

where dx_m and dx_n are the infinitesimal transformations of the parameterized curves on the surface, and the g_{mn} are the quantities that depend on those curves. The curvature of a surface is then determined from this fundamental quadratic form; this is the point of departure of Riemann's work. The local or infinitesimal approach favoured by Riemann allows for a generalization of Gauss' surfaces into the n-dimensional manifold with the metric groundform or fundamental metric tensor [7].

$$ds^2 = g_{\mu\nu}dx^{\mu}dx^{\nu} \quad (\text{for } \mu,\ \nu = 1,\ldots,\ n)$$

which controls the behaviour of the geodesics for a point and a direction in an n-dimensional manifold. Again it is this quadratic form or homogenous polynomial of the second degree that pervades Minkowski's geometry of numbers as well as his physical geometry or "physical axiomatics".

In the contemporary physics of string theory or M-theory (*à la* Witten), one could suppose that space and time are only illusions (not of the transcendental sort!). Quantum loop theory on the other hand is background-independent and presupposes only an invariant topology at the foundation of the physical universe. For General Relativity, the general covariance principle for all coordinate systems reduces to the invariance of the Minkowskian metric (asymptotically) in the gravitational field of local systems by the equivalence principle. Against the substantivalism of the space-time manifold and Einstein's hole argument, it suffices to identify events in spacetime with the intersection points of a Minkowskian point lattice in order to obtain a relational (Leibnizian) theory where diffeomorphism transformations swallow the Einsteinian hole or any background structure in an infinitesimal structure the internal logic of which is a geometry of numbers devoid of any physical ontology.

That is the end result of the import of arithmetical geometry in physical geometry from a foundational point of view. Einstein's own solution to the hole problem was to conclude that space-time locations do not have any physical meaning. As a matter of fact, the demise of substantivalism should mean the dismissal of Minkowski's *"Substanz"* and of its *"Postulat der absoluten Welt"*. The metric structure of the world does not match the matter fields and the whole problem (of the hole!) would simply vanish if one was ready to accept the idea that the metric field has no direct significance and that only a theory of gauge transformations could overcome

Einstein's difficulties. In 1919 Hermann Weyl introduced the concept of (local) gauge invariance in his unified field theory of gravitation and electromagnetism [6]. Weyl defended the view that the fundamental metric form is derived from the differential structure of the four-dimensional manifold and that the material content of the physical world is constrained by that structure, and not by the observational framework.

From Weyl's constructivist standpoint, Minkowski's absolute world view could translate into an *"imago mundi"*, a scientific image, as Weyl says, in the theoretical or symbolic construction of the world in physico-mathematical terms.

References

1. Gauthier, Y.: Internal Logic. Foundations of Mathematics from Kronecker to Hilbert. Kluwer (Synthese Library), Dordrecht/Boston/London (2002)
2. Gauthier, Y.: Hermann Weyl on Minkowskian Space-Time and Riemannian Geometry. Int. Studies Philos. Sci. **19**, 262–269 (2005)
3. Hilbert, D., von Neumann, J., Nordheim, L.: Über die Grundlagen der Quantenmechanik. Math. Ann. **98**, 1–30 (1928)
4. Hilbert, D.: Gesammelte Abhandlungen, 3 Bände. Chelsea, New York (1932)
5. Minkowski, H.: Gesammelte Abhandlungen. 2 Bände. Hrsg. v. D. Hilbert. Chelsea, New York (1967)
6. Weyl, H.: Raum, Zeit, Materie: Vorlesungen über allgemeine Relativitätstheorie. [s.n.], Berlin (1919)
7. Weil, H.: Philosophy of mathematics and natural science. Atheneum, New York (1960)

Chapter 11
The Mystical Formula and The Mystery of Khronos

Orfeu Bertolami

Abstract In 1908, Minkowski put forward the idea that invariance under what we call today the Lorentz group, $GL(1, 3, \mathbf{R})$, would be more meaningful in a four-dimensional space-time continuum. This suggestion implies that space and time are intertwined entities so that, kinematic and dynamical quantities can be expressed as vectors, or more generally by tensors, in the four-dimensional space-time. Minkowski also showed how causality should be structured in the four-dimensional vector space. The mathematical formulation proposed by Minkowski made its generalization to curved spaces quite natural, leaving the doors to the General Theory of Relativity and many other developments ajar.

Nevertheless, it is remarkable that this deceptively simple formulation eluded many researchers of space and time, and goes against our every day experience and perception, according to which space and and time are distinct entities. In this contribution, we discuss these contradictory views, analyze how they are seen in contemporary physics and comment on the challenges that space-time explorers face.

Keywords Space-time unification · Irreversibility · Arrows of time · Time in quantum gravity · Closed time-like curves · Cyclic time

11.1 The Mystical Formula

On the 21st of September 1908, at his address to the 80th Assembly of German Natural Scientists and Physicians in Cologne, Hermann Minkowski (1864–1909) presented his World Postulate, *Weltpostulat*, according to which the invariance under what we call today the Lorentz group $GL(1, 3, \mathbf{R})$, would be more meaningful in a four-dimensional space-time continuum [1]. This follows the very spirit of special relativity, in that the independence of the laws of physics on the velocity of inertial

O. Bertolami (✉)
Instituto Superior Técnico, Departamento de Física,
Av. Rovisco Pais 1, 1049-001, Lisboa, Portugal
e-mail: orfeu@cosmos.ist.utl.pt

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_11,

frames requires that space and time are indissoluble concepts and are related to each other by the maximum attainable particle velocity. In his own words:

"The views of space and time which I wish to lay down for 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."

To further stress the somewhat unusual nature of his proposal, Minkowski simply states that "no one can ever refer to a *Space* without its *Time* or to a *Time* without its *Space*" and actually, much more emphatically, "the essence of the World Postulate, which is pregnant with mathematical implications, could be dressed in the mystical formula:"

$$3 \times 10^5\, km = \sqrt{-1}\, sec. \,, \tag{11.1}$$

where the imaginary unit arises as Minkowski chooses to view space-time as strictly Euclidean, in its signature and in its lack of curvature. Thus, an event in the space-time continuum should be referred to as a world-point, "Welt-punkt", while its evolution in space-time continuum through a world-line, "Welt-linie". Thus, according to Minkowski, "all the world presents itself quite explicitly through world-lines" to the point that in his opinion, "physical laws would find their most comprehensive formulation through the reciprocal relationships of world-lines".

It is clear that the suggestion of a space-time continuum represents a further unification of concepts in physics (for a discussion on the inconsistency of a three-dimensional world see [2]). Indeed, special relativity allowed for a unified description of the laws of physics as well as for an unique formulation of mass, energy and momenta, thanks to the invariance of the maximum attainable particle velocity, c_{ST}, the Relativity Principle. It is important to remember that from the Relativity Principle, in any physical setting, distances can be measured by clocks and mirrors. Furthermore, an immediate implication of this order of ideas is that if space is isotropic, then any attempt to measure the time difference in the time of travel of light of equal distance paths would yield, irrespective of the direction, a null result. The most recent experimental attempts to measure deviations from this null outcome have shown that it holds up to a few parts in 10^{-9} [3]. Indirect experiments involving, for instance, ultra high-energy cosmic rays yield even more impressive limits, actually 1.7×10^{-25} (see e.g., [4] and references therein).

At this point, two comments are in order. The first one refers to the fact that the identification of the maximum attainable particle velocity with the speed of light, c, is only possible because, up to the current experimental precision, the photon mass vanishes and electromagnetism is an exact abelian gauge theory. If this were not the case, these two velocities could not be the same. Of course, historically, these two velocities were not initially distinguished and current bounds on the photon mass are compatible with this identification (see for instance [5]). Naturally, the same can be said about the identification of the speed of light and the velocity of propagation of gravitational waves in vacuum [6]. Thus, given the present bounds on the the photon (and also the graviton) mass we shall simply identify c_{ST} with c.

The second comment refers to the fact that the Standard Model (SM) vacuum being non-trivial, might not respect Lorentz invariance and hence correspond to a sort of preferred frame [7]. Alternatively, one can consider that only particle Lorentz invariance is physically meaningful, a perspective which allows for an extension of the SM compatible with the spontaneous breaking of Lorentz invariance, without implying the existence of a preferred frame [8]. Of course, this possibility would lead to distinct experimental signatures, which so far have not been observed (see e.g., [9] for comprehensive discussions).

So according to Minkowski, the motion of particles in the space-time continuum correspond to lines, *world lines*, from a given point in space-time where the original *event* took place. Past and future and hence causality are referred to this original event. In terms of the metric

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2 \ , \tag{11.2}$$

space-time admits *light-like* world lines, for which $ds^2 = 0$, *time-like* interval, for which $ds^2 < 0$, and a *space-like* interval, for which $ds^2 > 0$. Thus, in space-time diagrams, where time is depicted in the vertical axis and space in the horizontal one, light travels in the cone, the *light-cone* or null curves described, for a given time interval Δt, by $c\Delta t = \pm\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}$. Events within this cone are time-like and given that observers move with a relative velocity $v < c$, word-lines within this cone connect events in the past or in the future of each other, whether they precede or succeed each other. An event, say A, outside the light-cone cannot influence or be influenced by any other event separated by A by a space-like interval.

Clearly, these relationships have an absolute and global nature given the independence of the velocity of light on the velocity of the frame of reference; however, special relativity makes the concept of an absolute simultaneity impossible and thus the idea of an universal present. Of course, as already mentioned, this together with the fact that time flows at different rates for different observers and that likewise, the perception of space is also tied up with the relative motion of different observers, drives one away from Newton's (1643–1727) conception of absolute time, defined in the first book of his *Principia Mathematica* in 1687: "Absolute, true, and mathematical time, in and of itself and of its own nature, without reference to anything external, flows uniformly and by another name is called duration".

The insightful formulation of Minkowski allowed for a straightforward generalization and that was the path followed by Einstein (1879–1955) from 1907 onward after realizing that Newtonian gravity did not fit within the framework of special relativity. Later on, in collaboration with his friend and Zurich's Technical University colleague, Marcel Grossmann (1878–1936), Einstein wrote a seminal paper, albeit not quite consistent, in 1913, where it was clearly spelled out that Riemannian geometry was actually the most general and natural setting for physics (see, for instance, [10] for a detailed account). Of course, these developments relied strongly on the nineteenth century work of Lobatchevski (1793–1856), Bolyai (1802–1866), Riemann (1826–1846) and Gauss (1777–1855), who have shown that flat spaces are a particular case of a much wider class of spaces with non-vanishing curvature. This was indeed a quite new idea, even though, space and time of day to day affairs, were still regarded as *a priori* concepts that preceded all experience and were independent of any physical phenomena in the Newtonian (and Kantian) sense. But, of course, general relativity revolutionized this view showing that space and time are actually associated with a given energy-matter distribution, so that the Newtonian perspective was, at best, just an approximation to the inner nature of space and time.

Actually, the notion of world-lines in general relativity is basically similar to the one in special relativity, with the difference that in the former, space-time can be curved. The dynamics of the metric is determined by the Einstein field equations and depends on the energy-mass distribution in space-time. As before, the metric defines light-like (null), space-like and time-like curves. Also, in general relativity, world lines are time-like curves in space-time, where time-like curves fall within the light-cone. However, a light-cone is not necessarily inclined at 45° to the time axis, if one adopts the unit system where $c = 1$. However, this is an artifact of the chosen coordinate system, and reflects the coordinate freedom, the very essential diffeomorphism invariance of general relativity. Any time-like curve admits a co-moving observer whose "time axis" corresponds to that curve, and, since no observer is privileged, we can always find a local coordinate system in which light-cones are inclined at 45° to the time axis. Furthermore, the world-lines of free-falling particles or objects, such as the ones of planets around the Sun or of an astronaut in space, are minimal length curves, the so-called geodesics.

However, one should keep in mind that general relativity contains geometries which defy the very core of physical reasoning. Indeed, the existence of *closed time-like curves* contradicts the essential features of causality and chronology. Moreover, singularities, unavoidable in general relativity, once closed time-like curves are absent and geometry is set by well-behaved matter-energy configurations, imply that geodesics cannot exist in the whole space-time.

The referred condition on matter-energy is fairly specific as in the Hawking-Penrose singularity theorems and is tied up with the physical nature of a manifold [11]. A Lorentzian manifold (M, g) is said to be physically well-behaved if it satisfies the *strong energy condition*:

$$R_{\mu\nu} V^{\mu} V^{\nu} \geq 0 \,, \tag{11.3}$$

for any time-like vector field, V^{μ}. From Einstein's equations this statement is equivalent, for more than two d-spatial dimensions, to the condition on the energy-momentum tensor and its trace, T,

$$T_{\mu\nu} V^{\mu} V^{\nu} \geq \frac{T}{d-1} V_{\mu} V^{\mu} \,. \tag{11.4}$$

This condition is fulfilled by spaces dominated by the vacuum with a positive cosmological constant ($\Lambda \geq 0$), and by a perfect fluid if $\rho + 3p \geq 0$, where ρ and p correspond to the energy density and isotropic pressure, respectively.

Of course, the fundamental assumption here is the connection between physical spaces with mathematical spaces that satisfy the Einstein field equations. A generic mathematical space is fundamentally described by a d-dimensional differentiable manifold M endowed with a symmetric, non-degenerate second-rank tensor, the metric, g. A manifold under these specifications is said to a pseudo-Riemannian manifold, (M, g), as it has a Lorentzian signature $(-, +, ..., +)$ - it is Riemannian if it has signature $(+, ..., +)$. A differentiable manifold admits a Lorentzian signature if it is noncompact or has a vanishing Euler characteristic. The Italian mathematician Tullio Levi-Civita (1873–1941), showed a well known theorem according to which a pseudo-Riemannian manifold has a unique symmetric affine connection compatible with the metric, being hence equipped with geodesics.

The spaces relevant to physics correspond to solutions of the Einstein equations with a cosmological constant,[1] Λ:

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{11.5}$$

where $R_{\mu\nu}$ is the Ricci curvature of M, R its trace, G is Newton's constant and $T_{\mu\nu}$ is the energy-momentum tensor of matter in (M, g).

Thus, through Minkowski unification, it was possible to intimately relate physics back to geometry, a connection quite dear to Galileo (1564–1642) and Newton, but somewhat lost in the nineteenth century physics. In a previous text by this author [21], this methodology was referred to as "Cézanne's principle", given the suggestive connection it has with the writings of the French painter who first reflected on the then new cubist revolution. According to Cézanne (1839–1906), the essence of the new movement was the description of nature through purely geometrical forms. One could further mention that the relativity revolution is seen by many thinkers to be somewhat similar to the one that took place in modern art through movements like futurism, cubism, and other "isms" which have shown to be possible to depict in a single plane various points of view, as well as the mutation of reality through superimposing images. In this way, time was introduced into the arts that were traditionally associated with space, such as sculpture, architecture and painting, in opposition to the ones associated with time, music and literature.

The fundamental insight of Minkowski allowed for the generalization of the special theory of relativity and its application for understanding the inner secrets of the microscopical world through the development of quantum field theory. As far as experimental evidence allows us to unravel, quantum field theory methods are consistent down to about 10^{-18} m. On the largest known scales, general relativity, where the most general space-time continuum is actually curved, cosmological evidence seems to fit the so-called cosmological standard model up to the horizon size, i.e. distances up to 10^{26} m (see, e.g., [13, 14]). This is an impressive vindication of Minkowski's World Postulate.

---

[1] The "natural" system of units is used: $c = \hbar = k_{-B} = 1$.

11.2 The Mystery of Kronos

If at the physical and conceptual level, one could assume that most of the merit of the Minkowski unification is due to the radically different view implied by special relativity a few years earlier thanks to the work of Einstein, one should realize that from the historical and philosophical standpoint, the proposal of Minkowski is a remarkable culmination of more than 2,000 years of research about the nature of space and time. In what follows we shall discuss some of the most conspicuous philosophical ideas about the space and time problematic. Our main sources for the ensuing discussion are [15–22].

The very first manifestations of articulate rational thinking about the origin of the world, myths of creation, often regarded space and time as inseparable, as the original creation act gave birth to both space and time - actually, likewise the modern theory of the Big-Bang. In ancient Hellenic period, space and time were seen as two essential features of reality, but in many instances, regarded as distinct entities.

Indeed, at first, space and time seem to be quite different. Space can be freely experienced as one can move in any direction without restriction. Time however, has a well defined direction. Past and future are clearly distinct as our action can affect only the latter. We have memory, but not precognition. Matter, organic or otherwise, tends to decay rather than organize itself spontaneously. There seems to exist at least three distinct spatial dimensions[2] while there is only a single time dimension. Actually, the fact that in many physical theories the time dimension is just a parameter turns it into an "invisible dimension" [26]. Notice that if the number of time dimensions is greater than 1, one expects all type of complications as, on quite general grounds, the Partial Differential Equations that describe the physical phenomena are ultra-hyperbolic, which leads to unpredictability, or in weird "backward causality" (see e.g., [12] for a discussion).

Let us resume the philosophical discussion. Space has always been regarded as the arena of all manifestations of nature. Everything lies in space and the intrinsic and fundamental relationships between the most basic elements of *everything* could be decomposed into points, straight lines and geometrical figures in two or three dimensions and whose properties were monumentally described by Euclid's

[2] The Finnish physicist Gunnar Nordström (1881–1923) was the first to speculate in 1909, that space-time could very well have more than four dimensions. A concrete realization of this idea was put forward by Theodor Kaluza (1885–1954) in 1919 and Oskar Klein (1894–1977) in 1925, who showed that an unified theory of gravity and electromagnetism could be achieved through a five-dimensional (four-spatial and a single time dimension) version of General Relativity. These extra dimensions in order to have passed undetected can be either compact and very small or very large if the known fundamental interactions, excluding gravity, can test only three-spatial dimensions. In any case, the extension of the number of spatial dimensions has been widely considered in attempts to unify all known four interactions of nature. For instance, the requirement that supersymmetry is preserved in four dimensions, from the original ten-dimensional superstring theory, implies that six dimensions of the world are compact [23]. Connecting all string theories through S and T dualities suggests the existence of an encompassing theory, M-theory, and that space-time is 11-dimensional [24].

(ca. 330–275 BC) geometry. These relationships would in turn reveal the intrinsic properties of space itself. Of course, the fundamental role of space in the Hellenic philosophical thinking was more than evident on the speculative thinking of the pre-Socratic Zeno (495–435 BC) and Pythagoras (ca. 569–500 BC) and many others after them. For instance, for Plato (428–349 BC), "God ever geometrizes". For Aristotle (384–322 BC), the "geometrical" method and proof was the intellectual reasoning model that should be used in natural sciences, ethics, metaphysics and so on.

There were however, instances where philosophical thinking hinted at a hidden connection between space and time. For instance, for Zeno, the Dichotomy, the Achilles - tortoise, and the Arrow paradoxes stressed the fundamental difficulty in reconciling motion, that is dislocation in space, actually along a straight line, with the concepts of continuity and divisibility. In his Stadium paradox, Zeno considers three rows of bodies lying on a line and how the opposite relative motion of two of the rows "proof that half the time may be equal to double the time" [28]. Of course, these puzzles reflect the immaturity of the mathematical thinking at Zeno's time concerning the infinitesimal. But, it is suggestive that Zeno was already seeing that paradoxes in space and time were related in the real physical world through motion. A contemporary physicist could not fail to see that the divisibility process considered by Zeno could not go on indefinitely, as the fundamental limitation of quantum effects on the fabric of space-time would arise at Planck length, 10^{-35} m (or equivalently Planck time level, 10^{-44} s), the length where the Schwarzschild radius of a particle equals its Compton wavelength.[3]

For Pythagoras, who was the first to understand that above the application of mathematical tomb rules stood the *proof* of the fundamentals behind the rules, one could argue that the association of mathematics with music implied an inner connection between geometry and *tempo*, the very essence of music, time.

Actually in the Hellenic mythology, more specifically in the Orphic cosmogonies, time had a particularly interesting standing. Khronos, the primeval god of time, emerged as a self-formed divinity at the beginning of creation [29]:

"Originally there was Hydros (Water) and Mud, from which Ge (Earth) formed solidified ... The third principle after the Hydros and Ge was engendered by these, and was a Drakon (Serpent) with extra heads of a bull and a lion and a god's countenance in the middle; it had wings upon its shoulders, and its name was Khronos (Unaging Time) and also Herakles. United with it was Ananke (Inevitability, Compulsion), being of the same nature, her arms extended throughout the universe and touching its extremities ..."

Thus, Khronos and Ananke encircled the cosmos from the time of creation, and their passage drives the circling of heaven and the eternal flow of time.

---

[3] Actually, in some quantum gravity approaches, as for instance in loop quantum gravity, space-time is suggested to have, at its minutest scale, presumably the Planck scale, $L_P \simeq 10^{-35}$ m, a discrete structure [25]. In superstring/M-theory, the space-time continuum is an emergent property that arises from the ground state excitations of closed strings, one of the fundamental objects of the theory.

From myth to rational thinking, time was insightfully dissected by Heraclitus (ca. 535–475 BC) and by Aristotle (384–322 BC). Heraclitus understood the world as a unit resulting from diversity in eternal transformation. Time is what allows events to occur as a result of a web of antinomies. The formulas:" You cannot step twice in the same river; for fresh waters are ever flowing in upon you" and "The sun is new every day", capture the powerful idea that "all things are flowing".

For Aristotle, "time is a measure of motion according to the preceding and the succeeding", time is associated to the evolution, change in quantity and/or quality, of all occurrences in nature. For the most influential disciple of Plato, time is intimately related to motion, and with the counting process that is, with numbers. Aristotle's view leads to an operational connection between time and any material system that can be used as a standard for measuring the passage of time: clepsydras, sand clocks, sun clocks, pendulums and clocks. Since ancient times, the motion of earth around the sun has been the measure of day to day activities. Thus, in essence, our closest connection to time is the very one put forward by Aristotle more than 2,000 years ago. This is irrespective of any technological development, whether we use the regularity of the astronomical motions in the solar system (which, of course, cannot be exact due to the complexity of the physics behind these motions) or atomic clocks, or even binary pulsars, possibly the most precise clocks in the universe. In practical terms, one defines the second, the fundamental unit of time, as $1/86{,}400$ of the duration of the average solar day, or $9, 192, 631, 770$ periods of transition the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom.[4]

The average solar motion is defined in terms the of the idealized uniform motion of the sun along the celestial equator. The difference between this idealized motion and the real motion is called "equation of time".

Coming back to the philosophic discussion, the vision of time by Saint Augustine (354–430) is remarkable in its modernity. He believed that the origin of time was the creation of the world. This was a fundamental precaution since God's eternity made its identification with time impossible. In his *Confessions* he expresses his view that the notion of present is the most fundamental feature of time:
"The present of things past is memory. The present of things at present is perception. The present of things in the future is expectation."

Another fundamental aspect of Saint Augustine's view of time is his rejection of the doctrine of cyclic history. In many ancient civilizations, Sumer, Babylon, Indian, Mayan, the regular patterns of tides, seasons and the cyclic motion of heavenly bodies entailed from the fact that time itself was circular. Day is followed by night, night by day, summer follows winter, winter by summer, old moon follows new, new one by old one, and so why not history itself ?

The cyclic temporal pattern was a noticeable feature of Greek cosmology. The Stoics believed that on every instance after the planets returned to their exact relative positions as at the beginning of time, the whole cosmos would become renewed.

[4] This definition concerns a cesium atom at rest at a temperature of $0K$, such that the ground state is defined at zero magnetic field.

Nemesius, Bishop of Emesa says at the fifth century: "Socrates and Plato and each individual man will live again ... And this restoration of the universe takes place not once, but again and again, indeed to all eternity without end". The Maya civilization of Central America believed that history would repeat itself every 260 years, a period of fundamental importance in their calendar. The ancient Indians (Hindus, Budhists, Jains) extended the idea of a Great Year, a full cycle, into a hierarchy of Great Years. The destruction and re-creation of individuals and creatures occurred in a day of Brahma. A day of Brahma lasted 4,000 million years. The elements themselves and all forms will undergo a dissolution into Pure Spirit, which then incarnates itself back into matter every lifetime of Brahma, that is, about 311×10^{12} years. The lifetime of Brahma is the longest cycle and is repeated *ad infinitum* (see, e.g. in [30] and references therein).

The notion of cyclic time was thus regarded more comfortable, as a time arrow would mean instability, and inevitably irreversible change. The *myth of eternal return* was a central idea of many ancient civilizations, exceptions being the Hebrews and the Zoroastrian Persians. Thus, it was through the Judaeo-Christian tradition and most definitely through Saint Augustine that linear and irreversible time got established in the Western culture. According to Saint Augustine, Christ's death and Crucification was an unique event, and from then on the cultural prominence of the Roman Christian Church took charge of "spreading the word" about the linear nature of time and history. But even before that it was considered heresy to claim otherwise and transgressors were being punished in an exemplary way. Despite that, defenders of the doctrine of circular time, sometimes referred to as the *annulars*, were not short of conviction. In the third century, Euforbo, a presumed member of the *annulars* sect, while burnt at the stake, is alleged to have screamed: "This happened already and will happen again. You do not light a pyre, you light a labyrinth of fire. If all the pyres that I have been were put together they would not fit on earth and would blind the angels. I have said that many times" [31].

In this context, it is particularly interesting to remark that Spinoza (1632–1677), who believed that everything could be attributed to a manifestation of God's inscrutable nature, and as such occurred by absolute logical necessity, defended the idea that time was unreal, and hence all emotions associated to an event in the future or in the past are contrary to reason. For this philosopher, whose ethics should follow as in Euclid's geometry from definitions, axioms and theorems, it should be understood that "in so far as the mind conceives a thing under the dictate of reason, it is affected equally, whether the idea be of a thing present, past or future".

The existence of physical time was also doubted by the eleventh century, Persian philosopher Avicenna, who argued that time exists only in the mind due to memory and expectation. It is remarkable that similar views were expressed by Einstein himself. Indeed, in a letter to his life long friend, Michelle Besso, he writes (see, e.g. [26, 27]:

"There is no irreversibility in the basic laws of physics. You have to accept the idea that subjective time with its emphasis on the now has no objective meaning".

Later, on the occasion of Besso's passing away, in a letter addressed to his widow and son, he says:

"Michelle has preceded me a little in leaving this strange world. This is not important. For us who are convinced physicists, the distinction between past, present and future is only an illusion, however persistent".

Also worth mentioning is the view of time of one of the most brilliant opponents of Newton, Gottfried Leibniz (1646–1716). He argued that time cannot be an entity existing independently of actual events. For Leibniz absolute space does not exist. Space is the relative configuration of bodies that exist simultaneously. Thus, time is the succession of instantaneous configurations, and not a flux independent of the bodies and their motion. It follows that as time concerns a chronology of events, in a universe where nothing happens, there is no time. This disagreement with the very basis of Newtonian mechanics lead Leibniz to suggest that mechanics should be built strictly in terms of observed elements. This view was shared by the science philosopher Ernest Mach (1838–1916) and Heinrich Rudolph Hertz (1857–1894), who actually developed a "relational" mechanics based on the ideas of Leibniz. It is well known that Einstein was particularly impressed by Leibniz's ideas on space and time and these inspired him when constructing the general theory of relativity, and also played an important role in his life long rejection of quantum mechanics.

Causation, and thus time ordering, is according to Hume (1711–1776) the basis of human understanding. He warns that: "We ought not to receive as reasoning any of the observations we make concerning identity, and the relations of time and place; since in none of them the mind can go beyond what is immediately present to the senses ... Causation is different in that it takes us beyond impressions of our senses, and informs us of unperceived existences." It is arguable whether on purely philosophical terms Hume's doctrine stands on its own; however the decomposition of human perception down to the physiology of nervous tissues, down to its chemistry and then primarily the causal character of the physical laws, render Hume's proposition quite plausible.

In the *Critic of Pure Reason*, first published in 1781, Immanuel Kant (1724–1804) advances ideas about space, time and actually the main metaphysical problems of his time that turned out to be particularly influential. For Hume, the law of causality is not "analytic", that is, a proposition in which the predicate is part of the subject, such as for instance, an "equilateral triangle is a triangle". Kant agreed in that causation was a crucial starting point, however for him this law is synthetic and known *a priori*. A "synthetic" proposition is the one that is not analytic. All propositions that we know only through experience are synthetic. An "empirical" proposition is one which we cannot know except through the sense-perception, either our own or that of someone's testimony. So are the facts of history and geography, as well as the laws of science in so far as our knowledge of their truth depends on observational data. An a "priori" proposition, on the other hand, although susceptible of elucidation by experience, has, after inspection, a basis other than experience. All the propositions of pure mathematics are *a priori*. Kant then poses the question: How are synthetic judgments a priori possible ? His solution can be expressed in the following way: The outer world can only excite our senses, but it is our own mental apparatus that orders our sensations in space and time, providing in this way the means through which we understand experience. Space and time are thus subjective, they are part of our apparatus of perception; however, precisely because of their a priori nature, whatever we experience will exhibit the features that can be dealt with through geometry and the science of time. Space and time, Kant argues, are not concepts; they are means of intuition, forms of viewing or looking at the world.

Of some interest to physics[5] is also the part of the *Critic of Pure Reason* which deals with the fallacies, the "antinomies", mutually contradictory propositions which can be both proved to be true. They that arise from applying space and time or the categories (things in themselves) to what cannot be experienced. Kant discusses four of such antinomies, each consisting of thesis and antithesis. The first states: "The world has a beginning in time, and is also limited as regards to space". The antithesis says: "The world has no beginning, and no limits in space; it is infinite in regard to both time and space." The second antinomy proofs that every composite substance is both, and is not, made up of simple parts. The third antinomy states that there are two kinds of causality, one associated to the laws of nature, the other concerning that of freedom. The antithesis maintains that there is only the causality related to the laws of nature. Finally, the fourth antinomy shows that there is, and there is not, an absolute necessary Being. In a subsequent section, Kant destroys all purely intellectual proofs of the existence of God, even though he clarifies that he has other reasons for believing in God.

The antinomies have greatly influenced another important German philosopher, George Friedrich Hegel (1770–1831). Hegel believed in the unreality of the separateness, whether atoms or souls. The world is not a collection of self-subsistence units; nothing, Hegel held, is ultimately and completely real except the whole. Related to this is his disbelief in the reality of space and time as such, as these, if taken as completely real, involve separateness and multiplicity, which he regards as an illusion or as a mystic insight. Wholeness is the reality and this is rational as the rational is real. The engine of his metaphysical view of the world was dialectic: a thesis, antithesis and synthesis which sets consistency with the whole. His dialectic method applied to history in his *Philosophy of History*, could arguably give unit, and meaning to revolutions and movements of human affairs at the level of ideological currents of thought. Atributable to Karl Marx (1818–1883) is the theory that actually, the ultimate cause of human affairs moves dialectically, due to the clash of conflicting means of economical production. It is interesting that Marx regarded his insight about the development of human society as being analogous to the teleological evolution of species, the engine and clock of biological change, as first proposed by Darwin (1809–1882) and Alfred Russel Wallace (1823–1913) in 1858.

For Henri Bergson (1859–1941), intelligence and intellect can only form a clear idea of discontinuity and immobility, being therefore unable to understand life and to think about evolution. Intellect tends to represent *becoming* as a series of states.

[5] One should keep in mind that in 1755 Kant anticipated, in his *General Natural History and Theory of the Heavens*, how from Newton's mechanics one could explain the origin of the solar system. This nebular hypothesis was actually made mathematically plausible by Laplace's (1749–1827) many decades later.

Geometry and logic, the typical products of intelligence, are strictly applicable to solid bodies, but to everything else reasoning must be checked by common sense. Actually, Bergson believes that the genesis of intelligence and the origin of material bodies are correlative and have been developed by reciprocal adaptation. For him, the growth of matter and intellect are simultaneous. Intellect is the power of seeing things as separate and matter is what is separated into distinct things. However, in reality, there are no separate solid things, only a continuous stream of becoming. Becoming being an ascendant movement that leads to life, or a descendant movement leading to matter.

For Bergson, space and time are profoundly dissimilar. The intellect is associated with space, while instinct and intuition are connected with time. Space, the characteristic of matter, arises from a dissection of the flux which is really illusory, although useful in practice. Time, on the contrary, is the essential feature of life or mind. "Wherever anything lives, there is, open somewhere a register in which time is being inscribed". But time here is not a "mathematical" time, an homogeneous assemblage of mutually external instants. Mathematical time, according to Bergson, is actually a form of space; the time which is the essence of life is what he refers to as *duration*. "Pure duration is the form which our conscious states assume when our ego lets itself live, when it refrains from separating its present state from its former states". Duration unites past and present into an organic whole, where there is mutual entanglement and succession with distinction.

We could not draw to an end this brief account on the philosophical thinking about the nature of space and time without a reference to the images that often arise in poetry and literature, where time in particular, is quite often insightfully evoked. Indeed, from the Rubáiyát of Omar Khayyám (1048–1131) to "La recherche du temps perdu" of Marcel Proust (1871–1922), from William Blake (1757–1827) to contemporary authors such as Imre Kertész and Paul Auster, time and memory are central themes in the literary context. One often finds quite profound glimpses on the nature of time. In "Du côte de chez Swann", Proust says:
"The past is hidden beyond the reach of intellect, in some material object (in the sensation that the object will give us). And as for that object, it depends on chance whether we come upon it before we ourselves die."

Actually, this fundamental impression which allows us to expand our imagination and build theories based on the discovery of these "time capsules", a rock containing a fossil, the light of a distant star, the cosmic microwave background radiation or an ancient picture.

Time, the cycles of life and the hope of prevalence as put forth by Shakespeare (1564–1616) in his LX sonnet [32]:

Like as the waves make towards the pebbled shore,
So do our minutes hasten to their end;
Each changing place with that which goes before,
In sequent toil all forwards do contend.
Nativity, once in the main of light,
Crawls to maturity, wherewith being crown'd,
Crooked eclipses 'gainst his glory flight,

And Time, that gave, doth now his gift confound.
Time doth transfix the flourish set on youth,
And delves the parallels in beauty's brow;
Feeds on the rarities of nature's truth,
And nothing stands but for his scythe to mow:
  And yet to times in hope my verse shall stand,
  Praising thy worth, despite his cruel hand.

11.3 Arrows of Time

By the second half of nineteenth century, the development of the kinetic theory of matter by Maxwell (1831–1879), Clausius (1822–1888) and Boltzmann (1844–1906) revived once again the discussion on the dichotomy between the linear evolution of time and the eternal recurrence of motion.

The idea of a cyclic time and of an eternal return was recovered in philosophy by Herbert Spencer (1820–1903) and Friedrich Nietzsche (1844–1900) about the same time that Poincaré (1854–1912) showed his well known recurrence theorem. For sure, their "proofs" cannot be considered rigorous by the standards of physics and mathematics; however, interestingly, the "proof" of Nietzsche contains elements which can be regarded as relevant for any discussion of the issue, such as a finite number of states, finite energy, no creation of the universe and chance-like evolution. In his *Dialectic of Nature*, the philosopher and revolutionary politician, companion and co-author with Karl Marx of the *Communist Manifesto*, Friedrich Engels (1820–1895) wrote in 1879:
" ... an eternal and successive repetition of worlds in an infinite time is the only logical conclusion of the coexistence of countless worlds in an infinite space. ... It is in an eternal cycle that matter moves itself."

Let us now turn to the physical discussion. Newton's equations have no intrinsic time direction, being invariant under time reversal; however, Poincaré showed in 1890, in the context of classical mechanics, a quite general recurrence theorem, according to which any isolated system, which includes the universe itself, would return to its initial state given a sufficiently long time interval.

Poincaré's theorem is proved to be valid in any space X on which there exists a one parameter map T_i from sets $[U]$ and a measure μ on X such that: (i) $\mu(X) = 1$ and (ii) $\mu(T_{t_0}(U)) = \mu(T_{t_0+t}(U))$ for any subset of X and any t_0 and t. In classical mechanics, condition (i) is ensured by demanding that space X is the phase space of a finite energy system in a finite box. If μ is the distribution or density function, ρ, in phase space and T_t is the evolution operator of the mechanical system (associated with the Hamiltonian or Liouville operator), then condition (ii) follows from Liouville's theorem: $d\rho/dt = 0$. It thus follows that classical mechanics is inconsistent with the Second Principle of Thermodynamics.

Of course, the recurrence issue was a key concern to Boltzmann, who in the 1870s realized that deducing irreversibility, an arrow of time, from the mechanics of atoms was impossible without using averaging arguments. It was in the context of his efforts to understand the statistical equilibrium with the Liouville equation that he obtained in 1872 a time-asymmetric evolution equation, now referred to as Boltzmann equation, satisfied by a single-particle distribution function of a molecule in a diluted gas. From this he could construct a mathematical function, the so-called $\mathcal{H}$-function which is a strictly decreasing function of time. Identifying the $\mathcal{H}$-function with entropy with minus sign, Boltzmann could claim to have solved the irreversibility problem at molecular level.

However, in order to arrive at his result Boltzmann had to rely on the "molecular chaos hypothesis" (Stosszahlansatz), i.e., on the assumption that molecules about to collide are uncorrelated, but following the collision they are correlated as their trajectories are altered by the collision. Ernest Zermelo (1871–1953), young assistant of Planck (1858–1947) in Berlin, and Johann Josef Loschmidt (1821–1895), friend of Boltzmann, argued that the time-asymmetry obtained by Boltzmann was entirely due to the time-asymmetry of the molecular chaos assumption. Twenty years later, Zermelo attacked Boltzmann once again, now armed with Poincaré's recurrence theorem. Boltzmann attempted to save his case through a cosmological model. He suggested that as a whole the universe had no time direction, but rather individual regions could be time-asymmetric when through a large fluctuation from equilibrium it would yield a region of reduced entropy. These low entropy regions would evolve back to the most likely state of maximum entropy, and the process would repeat itself in agreement with Poincaré's theorem.

Having become clear that a finite system of particles would be recurrent and not irreversible in the long run, Planck considered whether irreversibility could emerge from a field theory such as electromagnetism. The point was to derive irreversibility from the interaction of a continuous field with a discrete set of particles. Starting to tackle the problem in 1897 in a series of papers, the work of Planck culminated with his discovery of the quantum theory of radiation in 1900. From Planck's arguments, Boltzmann remarked that as a field can be regarded as a system with an infinite number of degrees of freedom, and hence expected to be analogous to a mechanical system with an infinite number of molecules, an infinite Poincaré recurrence period and thus agreement with the observed irreversibility and the Second Principle would follow.

However, the persistent objections of influential opponents such as Ernest Mach and Friedrich Ostwald (1853–1932), led Boltzmann into depression and a first suicide attempt while at Leipzig before assuming his chair in Vienna in 1902. The intellectual isolation, as he was the sole survivor of the triumvirate of theoreticians along with Clausius and Maxwell, who had developed the kinetic theory of matter, and the continuous deterioration of his health led him to suicide and death at Duino, a seaside holiday resort on the Adriatico coast near Trieste, on the 5th September 1906. He was 62 years old. Boltzmann death is even more tragic when one realizes that it happened on the very eve of the vindication of his ideas.

However, the irreversibility problem has somehow resisted a straightforward answer. In 1907, the couple Ehrenfest, Paul Ehrenfest (1880–1933) and Tatyana Afanasyeva (1876–1964), (see, e.g. [33]), further developed Boltzmann's idea of averaging over a certain region Δ, of the phase space and showed that the averaged $\mathcal{H}$-function would remain strictly decreasing in the thermodynamical limit, after which Δ could be taken as small as compatible with the uncertainty principle.

In 1928, Pauli (1869–1958) considered the problem of transitions in the context of quantum mechanical perturbation theory and showed that consistency with the Second Principle of Thermodynamics would require a "master equation":

$$\frac{dp_i}{dt} = \sum_j (\omega_{ij} p_j - \omega_{ji} p_i) , \tag{11.6}$$

where ω_{ij} is the conditional probability per unit of time of the transition $j \to i$ and p_i is the probability of state i. Assuming the H-function to be given by

$$\mathcal{H} = \sum_i p_i \ln p_i , \tag{11.7}$$

it follows that $\frac{d\mathcal{H}}{dt} \leq 0$. This approach is quite suggestive as it indicates, as stressed by Boltzmann, that irreversible phenomena should be understood in the context of the theory that best describes microscopic physics.

More recently, Prigogine (1917–2003) and collaborators put forward a more radical approach, according to which irreversible behaviour should be already incorporated in the microscopic description (see [26] for a pedagogical discussion). In mathematical terms the problem amounts to turning time into an operator which does not commute with the Liouville operator, the commutator of the Hamiltonian with the density matrix. In physical terms, this proposal implies that the reversible trajectories cannot be used, leading to an entropy-like quantity which is a strictly increasing function of time.

But, if the problem of explaining the irreversible behaviour of all macroscopic systems from microphysics is already quite difficult, one should realize that there exists in nature quite a variety of phenomena whose behaviour indicate an immutable flow from past to present, from present to future. The term "arrow of time", already used in text, was coined by the British astrophysicist and cosmologist Arthur Eddington (1882–1944) [34] to characterize this evolutionary behaviour. Let us briefly describe these phenomena:

1. The already discussed time asymmetry inferred from the growth of entropy in irreversible and dissipative phenomena, as described by the Second Law of Thermodynamics.
2. Nonexistence of advanced electromagnetic radiation, coming from the infinite and converging to a source, even though solutions of this nature are legitimate solutions of the Maxwell's field equations.
3. The collapse of wave function of a quantum system during the measurement process and the irreversible emergence of the classical behaviour, even though the fundamental equations of quantum mechanics and statistical quantum mechanics, Schrödinger's and Von Neumann's equations, respectively, are invariant under time inversions for systems described by time-independent Hamiltonians (see e.g., [35] for a vivid discussion).

4. The exponential degradation in time of systems and the exponential growth of self-organized systems (given a sufficiently large supply of resources). In the development of self-organized systems, a particularly relevant role is played by complexity. The fascinating aspects of phenomena in this context has lead authors to refer to them as "creative evolution", "arrow of life", "physics of becoming" [17, 18, 26, 27]. In these discussions, the chaotic behaviour plays an important role given that complex systems are described by non-linear differential equations. This chaotic behaviour gives origin to an extremely rich spectrum of possibilities for describing self-organized systems as well as a paradoxically predictable randomness as chaotic branches are deterministic (see for instance [17, 36]).
5. The discovery of the CP-symmetry violation in the $K^0 - \bar{K}^0$ system implies, on account of the CPT-theorem, that the T-symmetry is also violated. This means that on a quite elementary level there exists an intrinsic irreversibility. The violation of the CP-symmetry and also of baryon number in an expanding universe are conditions from which the observed baryon asymmetry of the universe can be established (see, e.g., [37] and references therein). An alternative way to achieve the baryon asymmetry of universe is through the violation of the CPT-symmetry (see [38] and references therein).
6. Psychological time is clearly irreversible and historical. The past is recognizable, while the future is open.
7. The so-called gravito-thermal catastrophic behaviour [39] of systems bound gravitationally, implies, given their negative specific heat, that their entropy grows as they contract beyond limit. On the largest known scales, the expansion of the Universe, which is adiabatic, is a quite unique event, and as such, is conjectured to be the master arrow of time to which all others are subordinated.

11.4 Open Questions

Let us briefly discuss in this section a few problems concerning the nature of spacetime that remain unsolved. These include the issue of a putative correlation between the above described arrows of time, the problem of nonexistence of an explicit time in the canonical Hamiltonian formulation of quantum gravity, the question of solutions of general relativity and other gravity theories that exhibit closed time-like curves and whether the universe evolves after all in a cyclic way.

11.4.1 Are The Arrows of Time Correlated ?

The existence of systems, from which a time direction can be inferred, is not on its own very surprising, as it is in the core of all dissipative phenomena. One could argue that this property reflects, for instance, a particular choice of boundary conditions which constrain the state of the universe, rather than any restriction on its dynamics and evolution. However, this point of view cannot account for the rather striking fact that the known arrows of time seem all to point in the same time direction. In the following, we shall briefly overview some of the ideas put forward to relate the arrows of time. Extensive discussions can be found in [40–42].

In his book *The Direction of Time*, the philosopher Hans Reichenbach (1891–1953) [43] argued in a rather circular way that the arrow of time in all macroscopic phenomena has its origin in causality, which in turn should be the origin of the growth of entropy. In 1958, the cosmologist Thomas Gold (1920–2004) put forward the remarkable idea that all arrows of time should be subordinated to the expansion of the universe [44]. This speculation gave origin to demonstrations, although not quite entirely successful, that the propagation of the electromagnetic radiation was indeed related to the expansion of the universe [45, 46]. The problem is that the obtained solutions are somewhat puzzling. Indeed, it is found that: retarded radiation is found to be compatible only with a steady-state universe, while advanced radiation is found to be compatible only with evolutionary universes (expanding or contracting ones). For sure, these solutions indicate that the problem is more complex than admitted.

Inspired by the Thermodynamics of Black Holes, Penrose put forward the suggestion that the gravitational field should have an associated entropy which, in turn, should be related with an invariant combination involving the Weyl tensor [47]. Remarkably this suggestion allows for a consistent set up for cosmology of the Generalized Second Principle of Thermodynamics, which arises in black hole physics, and states that the Second Principle should apply to the sum of the entropy of matter with the one of the black hole [48, 49]. The main point of the proposal is that it circumvents the paradox of an universe whose initial state is a singularity or a black hole protected by a horizon, and hence with an initial entropy that exceeds by many decades of magnitude the entropy of the observed universe. Being highly homogeneous and isotropic, the initial state of the universe has necessarily a low entropy[6] as the Weyl tensor vanishes for homogeneous and isotropic geometries. The gravitational entropy will then increase as the Weyl tensor increases as the universe grows lumpier.

The growth of the total entropy can presumably account for the asymmetry of psychological time as in this way the branching of states and outcomes will occur into the future.

Let us close this subsection with some remarks on some recent ideas developed in the context of superstring/M-theory. These suggest a multiverse approach of the "landscape" of vacua of the theory (see, e.g. [12] and references therein), that is the googleplexus of about 10^{500} vacua [51], which are regarded as distinct universes, which asks for a selection criteria for the vacuum of our universe. Anthropic arguments [52] and quantum cosmological considerations [53] have been advanced for

[6] The low entropy of the highly "excited" and hot initial state was suggested to be analogous of systems with a negative temperature [50].

this vacuum selection as a meta-theory of initial conditions. These proposals are not consensual, but can be seen as a relevant contribution to a deeper understanding of the problem. Of course, one should keep in mind that non-perturbative aspects of string theory are still poorly known [54]. The multiverse perspective hints at the possibility that different universes may actually interact [55]. It is suggested that this interaction is regulated by a Curvature Principle and shown, in the context of a simplified model of two interacting universes, that the cosmological constant of one of the universes is driven toward a vanishingly small value. The core of this proposal is to set an action principle for the interaction using the curvature invariant $I_i = R^i_{\mu\nu\lambda\sigma} R_i^{\mu\nu\lambda\sigma}$, where $R^i_{\mu\nu\lambda\sigma}$ is the Riemann tensor of each universe. The suggested Curvature Principle also hints at a solution for the entropy paradox of the initial state of the universe [55]. For this, one considers the point of view of another universe, from which our universe can be perceived as if all its mass were concentrated in some point and hence $I = 48M^2r^{-6}$, where r is the universe horizon's radius and M its mass-using units where $G = \hbar = c = 1$. Hence, if the entropy scales with the volume, then $S \sim r^3 \sim I^{-1/2}$; if the entropy scales according to the holographic principle, suitable for AdS spaces [56, 57], then $S \sim r^2 \sim I^{-1/3}$. In either case, given that $I \sim \Lambda^2$ for the ground state, one obtains that $S \to 0$ in the early universe and, $S \to \infty$ when $\Lambda \to 0$. The latter corresponds to the universe at late time, which is consistent with the Generalized Second Principle of Thermodynamics.

Of course, a multiverse perspective, if taken to its most extreme versions, can lead to intricate problems concerning the relationship among the cosmic time of each universe and the "meta-time" of the whole network of universes. Only the future will tell us whether developments in this direction will be needed to further understand the physics of our universe.

11.4.2 Time in Quantum Gravity

Quantum gravity, the theory that presumably describes the behaviour of space-time at distances of the order of the Planck length is still largely unknown. The most developed programme to understand quantum gravity, superstring/M-theory, leads to a quite rich lore of ideas and concepts, but has not provided so far a satisfactory answer concerning for instance, the fundamental problem of smallness of the cosmological constant [58], and exhibits the vacuum selection problems discussed above, which seriously threaten the predictability power of the whole approach.

In order to understand the conceptual difficulties of the quantum gravity problem, let us see that from its very beginning, the quantization of gravity poses outstanding challenges to the well known and well tested methods of quantum field theory. Indeed, if one considers the metric, $g_{\mu\nu}(\boldsymbol{r}, t)$, a bosonic spin-two field and attempts its quantization through an equal-time commutation relation for the corresponding operator:

$$[\hat{g}_{\mu\nu}(\boldsymbol{r}, t), \hat{g}_{\mu\nu}(\boldsymbol{r}', t)] = 0 \,, \tag{11.8}$$

for $\boldsymbol{r} - \boldsymbol{r}'$ space-like, then one faces an indefinite problem: (i) In fact, in order to establish that $\boldsymbol{r} - \boldsymbol{r}'$ is space-like, one must specify the metric; (ii) being an operator relationship, it must hold for any state of the metric; (iii) without specifying the metric, causality is ill-defined.

These difficulties compel one to consider a canonical quantization programme based on Hamiltonian formalism (see, e.g. [59] and references therein). In this context, one splits space and time and selects foliations of space-time where the physical degrees of freedom of the metric are the space-like ones, $h_{ab} = {}^{(3)}g_{ab}$. The resulting Hamiltonian is a sum of constraints, one associated with invariance under time reparametrization, the others related with invariance under three dimensional diffeomorphisms. If one considers only Lorentzian geometries (a quite restrictive condition !), then only the first constraint is relevant. The solution of the classical constraint is given by:

$$H_0 = 0 \,, \tag{11.9}$$

where

$$H_0 = \sqrt{h}\left[h^{-1}\Pi_{ab}\Pi^{ab} - (3)_R\right] \,, \tag{11.10}$$

h being the determinant of the 3-metric h_{ab}, Π_{ab} the respective canonical conjugate momentum and ${}^{(3)}R$ the 3-curvature. Quantization follows by turning the momenta into operators for some operator ordering and applying the resulting Hamiltonian operator into a wave function, the wave function of the universe, $\Psi[h_{ab}]$:

$$\hat{H}_0\Psi[h_{ab}] = 0 \,. \tag{11.11}$$

This is the well known Wheeler-DeWitt equation.

In this context, the problem of time (see [60] for a detailed account) consists in not having a Schrödinger-type equation for the evolution of states, but instead, the constrained problem (11.11), where time is one of the variables within H_0. Of course, this does not mean that there is no evolution, but rather that there is no straightforward way of extracting a variable from the formalism that resembles the cosmic time one is used to in classical cosmology.

Solutions, although partial, include the semi-classical approach [61, 62], where time is identified with the scale factor or some function of it, once the metric starts behaving like a classical variable and the wave function of the universe admits a WKB approximation. In this instance, the Wheeler-DeWitt equation can be written, at least in the minisuperspace approximation, as the Hamilton-Jacobi equation for the action of the WKB approximation. Physically it implies that time is meaningful only after the metric becomes classical.

Another interesting idea is the so-called "Heraclitean time proposal" [63, 64]. This is based on a suggestion by Einstein [65] according to which the determinant of the metric might not be a dynamical quantity. In this theory, usually referred to as unimodular gravity, the cosmological constant arises as an integration constant and an "Heraclitean" time can be introduced as the classical Hamiltonian constraint assumes the form [63]:

$$H = \Lambda h^{1/2} , \tag{11.12}$$

and thus, for a given space-like hypersurface Σ, one can write

$$i\frac{\partial \Psi}{\partial t} = \int_{\Sigma} d^3x h^{-1/2} \hat{H}_0 \Psi = \hat{H}\Psi , \tag{11.13}$$

which has a Schrödinger-like form.

For sure, the problem of time in quantum gravity still remains an open question and the above approaches were presented only to exemplify some possible directions for future research.

11.4.3 Closed Time-Like Curves and Time Travel

As already mentioned, closed time-like curves arise as solutions of Einstein's field equations. These solutions include traversable wormholes [22, 66–68], warp drives [69, 70] and the Krasnikov tube [71]. One can argue that they are unphysical as they violate the energy conditions [72]. These solutions correspond to putative forms of time travel and most often bring a host of paradoxes of the ancestor's murder type. However, given that the murder of an ancestor by a time traveler should be logically inconsistent, one could ask whether there should exist global self-consistent conditions to exclude closed time-like curves. These conditions are referred to as *consistency constraints*. The most discussed of these consistency constraints are the Principle of Self-Consistency [73] and the Chronology Protection Conjecture [74].

The Principle of Self-Consistency states that events along a closed time-like curve are self-consistent, that is they influence each other, but in a self-consistent fashion. Of course, along a closed time-like curve the notion of past or future is ambiguous and the causal structure of usual space-times is meaningless. The self-consistent condition establishes that events in the future can influence events in the past, but cannot alter them. Hawking's Chronology Protection Conjecture is based on the experimental evidence that "we have not been invaded by hordes of tourists from the future" [74] from which it is then conjectured that the renormalized stress-energy tensor quantum expectation values diverge as they approach closed time-like curves. This divergent behaviour destroys the wormhole's structure before the Planck scale is attained. So far, no proof of this conjecture is available.

Thus, one sees that the reality of closed time-like curves may be contested on physical as well as on logical grounds. Nevertheless, these solutions are vivid examples of the wealth of structurally distinct solutions of general relativity and show how some classes of solutions may require a specific set of criteria to establish their physical reality.

11.4.4 A Cyclic Time ?

The general theory of relativity allows a for a global dynamical description of the physical space-time and for a relation with the history and evolution of the universe. The mathematical description of space-time admits a wide range of scenarios, which includes solutions with cyclic nature. Already in 1922 Alexander Friedmann (1888–1925), the first to study evolving cosmological solutions within general relativity, realized that cyclic scenarios existed among his solutions. These involved an expanding universe followed by a recollapse so that the universe's radius would eventually vanish from which a new expansion would ensue. Of course, strictly speaking these cycles are not mathematically admissible as they are disjoint by a singularity. In 1931, Richard Tolman (1881–1948) [75] showed that such discontinuity was unavoidable at the beginning and at the end of any isotropic and homogeneous closed geometries for a physically realistic energy-momentum tensor. Subsequently, he argued that the problem was actually due to the highly symmetric nature of the studied solutions and that in a physically realistic universe the discontinuity could very well disappear [76].

A cyclic or "phoenix" universe was regarded with sympathy by Einstein and George Gamow (1904–1968), who even coined the term "big squeeze" to denote the final state of collapse - nowadays the term "big crunch" is more used. Of course the issue of space-time singularities was not fully appreciated then; however, in the 1960s it was understood, through the Hawking-Penrose singularity theorems, that the conditions and the generality of the difficulty could not be overlooked and cosmologists had to accept the reality of the space-time singularities. Some relativists argued however, that quantum effects could play a role in the process of "bouncing" at very high densities completing in this way the cycle of a closed universe. John Wheeler, for instance, advocated that in the "bounce" physical constants would be recycled [77].

More recently, developments in string theory and the related dynamics of branes do open the possibility of reviving the idea of a cyclic universe. In the so-called "ekpirotic" model [78], one assumes, as a starting point, the existence of two three-dimensional parallel branes embedded in a higher dimensional space. Our universe corresponds to one of these branes. Quantum fluctuations in the other brane would lead to the creation of a third brane, which would be attracted to ours. The ensuing impact of the third brane into ours would trigger a release of energy, *ekpirosis* in Greek, giving origin to a proto-universe, whose subsequent expansion would have properties similar to the ones of a universe just emerging from the inflationary process. Thus, this collision process is quite similar to the Big-Bang itself. Of course, whether a universe emerging from the ekpirotic process fully resembles our universe, or whether it advantageously replaces the inflationary dynamics, whose most generic features are consistent with the latest observations of the cosmic microwave background radiation [79], is still a quite open question. It is interesting that these two competing models have a distinct behaviour in what concerns the production of primordial gravitational waves. The ekpirotic process tends not to produce too much gravitational waves, while some models of inflation do produce a considerably greater amount of gravitational radiation. The possibility of verifying the prediction of these models through the observation of gravitational waves is of course quite exciting.

The cyclic nature of the "ekpirotic" model arises from the fact that after several decades of thousands of millions of years after the brane collision, our universe will expand to the point where stars and galaxies will be all gone and there will be no radiation left. This void and cold brane will be very similar to the one before the Big-Bang. Conditions will then be favourable for the creation of a third brane from the other original brane and the whole process then repeats itself.[7]

11.5 Conclusions and Outlook

The unification of space and time proposed by Minkowski a century ago allowed for an elegant formulation of the special theory of relativity, and was the culmination of more than 2,000 years of philosophical and physical research on the nature of space and time. The space-time continuum is a basic foundational concept in physics, from elementary particle physics to cosmology. The vector space structure of the space-time continuum made the transition to the general theory of relativity smooth and quite logical once it was understood that, at cosmological scales, the space-time continuum was not an *a priori* concept, independent of the physical conditions. Furthermore, when analyzing the inner makings of matter, Minkowski's space-time formulation, together with quantum mechanics, made possible, through renormalizable quantum field theory, to stretch our knowledge down to scales of about 10^{-18} m. The research on the matching of general relativity with the quantum nature of matter is still in its infancy; however, we already understand that reconciling these two pillars of the twentieth century physics will require a whole new set of ideas, as it may happen, that we may have to give up concepts that were supposed to be the starting point of all the modeling of the universe, such as that space-time is a continuum and that the fundamental building blocks of matter are not point-like particles.

These assumptions lead to quite new realms for research and experimentation. They also pose us new technical and conceptual problems. These imply that the very principles upon which our theories of space-time were built so far, such as Lorentz invariance, CPT-symmetry, the commutative nature of the fundamental dynamical variables and so on, will have to be continuously scrutinised. Their breakdown may provide important insights into the nature of the new theories of space-time, matter and the universe. Of course, these new theories will have to match smoothly our current physical theories and explain the conditions for the emergence of the Minkowski space-time continuum as well as to set the boundaries of validity of

[7] The similarity with the Indian mythology is compelling. Each cycle is analogous to the "day of Brahma". The whole process resembles the "life of Brahma".

general relativity and the emergence of the classical features of gravity. The new theories will, like in the case of general relativity, pose questions of ontological nature and should set criteria for selecting, among the mathematically consistent solutions, the physical ones which have predictive power to explain our world. The most recent developments in the context of superstring/M-theory, the most studied quantum gravity programme, suggest that a multitude of universes is needed to explain the physics of our universe. This is a somewhat disappointing outcome for a theory that naturally unifies quantum mechanics and general relativity. However, this may only reflect the provisional state of our knowledge.

On the other hand, as we have discussed, the quest for the understanding of the ultimate nature of space-time, and the rather special role played by time in macrophysics and its various arrows is still largely unknown. If all arrows of time can be related with the expansion of the universe, or to some new curvature principle that properly accounts for the entropy of the gravitational field, a remarkable new unification could be achieved.

In any case, the quest for the ultimate theories about the nature of space and time have mesmerized human thought for more than 2,000 years. Till recently, the most insightful ideas sprang from philosophical speculation, however since the pioneering work of Einstein on relativity and the space-time unification proposed by Minkowski a century ago, physicists have taken the lead in this search. A hundred years after the proposal of Minkowski, mankind is about to embark on new expeditions to conquer new continents of knowledge through new scientific challenges which include the Large Hadron Collider to search for the nature of mass and new symmetries, and new space missions to study the polarization of the cosmic microwave background radiation and to directly detect gravitational waves. It is the hope of the whole scientific community that the outcome of these experiments will bring precious hints for the understanding of our universe.

References

1. Minkowski, H.: Raum und Zeit, 1908. Also in Lorentz, A.H., Einstein, A., Minkowski, e H. (eds.) O Princípio da Relatividade. Fundação Calouste Gulbenkian, Lisboa (1978)
2. Petkov, V.: Relativity, dimensionality, and the existence, In: Petkov, V. (ed.) Relativity and the dimensionality of the world, Springer Fundamental Theories of Physics 153. Springer, AA Dordrecht, The Netherlands (2007)
3. Müller, H. et al.: Phys. Rev. Lett. **91**, 020401 (2003); Wolf, P. et al.: Phys. Rev. Lett. **90**, 060402 (2003)
4. Bertolami, O., Carvalho, C.: Phys. Rev. D. **61**, 103002 (2000)
5. Bertolami, O., Mota, D.F.: Phys. Lett. B. **455**, 96 (1999)
6. Ellis, G.F.R., Uzan, J.P.: c is the speed of light, isn't it?, gr-qc/0305099
7. Consoli, M., Costanzo, E.: Is the physical vacuum a preferred frame?, arXiv:0709. 4101[hep-ph]
8. Colladay, D., Kostelecký, V.A.: Phys. Rev. D. **55**, 6760 (1997); **58**, 116002 (1998)
9. Kostelecký, A. (ed.): CPT and Lorentz Symmetry III, World Scientific, Singapore (2005); Bertolami, O.: Gen. Rel. Gravitation **34**, 707 (2002); Bertolami, O.: Lect. Notes Phys. **633**,

96 (2003) hep-ph/0301191; Mattingly, D.: Liv. Rev. Rel. **8**, 5 (2005), gr-qc/0502097; Lehnert, R.: "CPT- and Lorentz-symmetry breaking: a review", hep-ph/0611177
10. Pais, A.: Subtil é o Senhor, Vida e Pensamento de Albert Einstein. Gradiva, Lisboa (2004)
11. Hawking, S.W., Ellis, G.F.R.: Large scale structure of space-time. Cambridge University Press (1973)
12. Bertolami, O.: The adventures of Spacetime. In: Petkov, V. (ed.) Relativity and the dimensionality of the world. Springer Fundamental Theories of Physics 153 Springer, AA Dordrecht, The Netherlands (2007)
13. Will, C.: The confrontation between general relativity and experiment. gr-qc/0510072
14. Bertolami, O., Páramos, J., Turyshev, S.: General theory of relativity: will it survive the next decade ? gr-qc/0602016
15. Russell, B.: History of western philosophy. Counterpoint, London (1946)
16. Abbagnano, N.: História da Filosofia. Editorial Presença, Lisboa (1984)
17. Coveney, P., Highfield, R.: The arrow of time. Fawcett Columbine, New York (1990)
18. Des hommes de science aux prises avec le temps. Group de Matheron. Presse Polytechniques e Universitaires Romandes, Laussane (1992)
19. Ellis, G.F.R.: Physics in a real universe: time and space-time. In: Petkov, V. (ed.) Relativity and the dimensionality of the world. Springer Fundamental Theories of Physics 153 Springer, AA Dordrecht, The Netherlands (2007)
20. Bertolami, O.: The concept of time in physics. In: Proceedings of the 7th International Conference of Physics Students, Lisbon (1992)
21. Bertolami, O.: O Livro das Escolhas Cósmicas. Gradiva, Lisboa (2006)
22. Lobo, F.S.N.: Nature of time and causality in physics. arXiv:0710.0428 [gr-qc]
23. Candelas, P., Horowitz, G., Strominger, A., Witten, E.: Nucl. Phys. B **258**, 46 (1985)
24. Witten, E.: Nucl. Phys. B **443**, 85 (1995)
25. Ashtekar, A., Rovelli, C., Smolin, L.: Phys. Rev. Lett. **69** 237 (1992)
26. Prigogine, I.: From Being to Becoming. W.H. Freeman & Co. New York (1980)
27. Progine, I., Stengers, I: La Nouvelle Alliance. Gallimard, Paris (1979)
28. Bell, E.T.: Men of Mathematics. Simon and Schuster, New York (1965)
29. Orphica, Fragments, Greek Hymns C3rd B.C. - C2nd A.D.
30. Tipler, F.J.: Essays in general relativity, Festschrift for Taub, A. (ed.) Academic Press, New York (1980)
31. Borges, J.L.: El Aleph. Emecé Editores, Buenos Aires (1957)
32. Shakespeare, W.: The Illustrate Stratford Shakespeare. Chancellor Press, London (1982)
33. Huang, K.: Statistical Mechanics. John Wiley, New York (1966)
34. Eddington, A.: The nature of the physical world. Cambridge University Press, Cambridge (1928)
35. Penrose, R.: The Emperor's New Mind. Vintage, London (1990)
36. Gleick, J.: Chaos. Cardinal, London (1988)
37. Buchmüller, W.: Baryogenesis - 40 years later. arXiv:0710.5857[hep-ph]
38. Bertolami, O., Colladay, D., Kostelecký, V.A., Potting, R.: Phys. Lett. B. **395**, 178 (1997)
39. Lynden-Bell, D.: Mont. Not. Roy. Astr. Soc. **123**, 447 (1962)
40. Davies, P.C.W.: The physics of the time asymmetry. California University Press, Berkeley (1974)
41. Zeh, H.D.: The physical foundation of the direction of time. Heidelberg University Preprint (1988)
42. Mann, M.G., Hartle, J.B.: Time symmetry and asymmetry in quantum mechanics and quantum cosmology. University of California Santa Barbara Preprint (1991)
43. Reichenbach, H.: The Direction of Time. California University Press, Berkeley (1956)
44. Gold, T.: La Structure et L'Evolution de L'Universe. 11th International Solvay Congress. Edition Stoops, Brussels (1958)
45. Hogarth, J.E.: Proc. R. Soc. London A **267**, 365 (1962)
46. Hoyle, F., Narlikar, J.V.: Proc. R. Soc. London A **277**, 1 (1964)
47. Penrose, R.: In: Hawking, S., Israel, W. (eds.) General relativity: An Einstein Centenary Survey. Cambridge University Press, Cambridge (1979)

48. Penrose, R.: Il Nuovo Cimento, Num. Spec. **I 1**, 252 (1969)
49. Bekenstein, J.D.: Phys. Rev. D. **7**, 2333 (1973); **9**, 3292 (1974); Hawking, S.W.: Comm. Math. Phys. **43**, 199 (1975)
50. Bertolami, O.: Negative Temperatures and the arrow of time. (1985); unpublished
51. Bousso, R., Polchinski, J.: JHEP **0006**, 006 (2000)
52. Susskind, L.: The anthropic landscape of string theory. hep-th/0603249; Susskind, L.: The cosmic landscape: string theory and the illusion of intelligent design. Little, Brown, New York (2005)
53. Holman, R., Mersini-Houghton, L.: Why the Universe Started from a Low Entropy State. hep-th/0511102; Holman, R., Mersini-Houghton, L.: Why did the universe start from a low entropy state ?. hep-th/0512070
54. Polchinski, J.: The cosmological constant and the string landscape. hep-th/0603249
55. Bertolami, O.: A curvature principle for the interaction between universes. arXiv:0705.2325[gr-qc]; to appear in General Relativity and Gravitation
56. Fischler, W., Susskind, L.: Holography and cosmology. hep-th/9806039
57. Bousso, R.: JHEP **9906**, 028 (1999)
58. Witten, E.: The cosmological constant from the viewpoint of string theory. hep-ph/0002297.
59. Bertolami, O., Mourão, J.M.: Class. Quant. Grav. **8**, 1271 (1991)
60. Isham, C.: Canonical quantum gravity and the problem of time. gr-qc/9210011
61. DeWitt, B.S.: Phys. Rev. **160**, 1113 (1967)
62. Vilenkin, A.: Phys. Rev. D. **39**, 1116 (1989)
63. Unruh, W.: Phys. Rev. D **40**, 1048 (1989); Unruh, W., Wald, B.M.: Phys. Rev. D **40**, 2598 (1989)
64. Bertolami, O.: Int. J. Mod. Phys. D **4**, 97 (1995)
65. Einstein, A.: Sitz. Berl. Preuss. Akad. Wiss. (1919)
66. Visser, M.: Lorentzian wormholes: from Einstein to Hawking. American Institute of Physics, New York (1995)
67. Morris, M., Thorne, K.S.: Am. J. Phys., **56**, 395 (1998)
68. Lobo, F.S.N.: Phys. Rev. D **71**, 084011 (2005); **71**, 124022 (2005)
69. Alcubierre, M.: Class. Quant. Grav. **11**, L73 (1994)
70. Lobo, F.S.N., Visser, M.: Class. Quant. Grav. **21**, 5871 (2004)
71. Krasnikov, S.V.: Phys. Rev. D **57**, 4760 (1998)
72. Morris, M., Thorne, K.S., Yurtsever, U.: Phys. Rev. Lett. **61**, 1446 (1988); Deser, S., Jackiw, R., t'Hooft, G.: Phys. Rev. Lett. **68**, 267 (1992); Deser, S.: Class. Quant. Grav. **10**, S67 (1993)
73. Earman, J.: Bangs, crunches, whimpers and shrieks: singularities and acausaliites in relativistic Spacetimes. Oxford University Press, Oxford (1995)
74. Hawking, S.W.: Phys. Rev. D **56**, 4745 (1992)
75. Tolman, R.: Relativity, Thermodynamics and Cosmology. Oxford University Press, Oxford (1934)
76. Tolman, R.: Phys. Rev. **38**, 1758 (1931)
77. Misner, C.W., Thorne, K.S., Wheeler, J.A.: Gravitation. Freeman, San Francisco (1973)
78. Turok, N., Steinhard, P.: Beyond inflation: a cyclic universe scenario. hep-th/0403020
79. Spergel, D.N. et al.: [WMAP Collaboration], Astrophys. J. Suppl. **170**, 377 (2007)

Chapter 12
Physical Laws and Worldlines in Minkowski Spacetime

Vesselin Petkov

Abstract In his paper "Space and Time" a hundred years ago Minkowski gave us the adequate relativistic picture of the world. According to him what exists is an absolute four-dimensional world in which the ordinary physical bodies are worldlines. Minkowski conjectured that physical laws might find their most perfect expression as interrelations between these worldlines. The purpose of this paper is to examine further whether Minkowski's idea can be applied to different areas of physics. It is shown that not only does it work perfectly in classical physics and general relativity, but also provides a deeper understanding of some difficult questions (including the origin of inertia) and demonstrates that taking seriously the existence of worldlines inescapably leads to the concept of gravity as curvature of spacetime. It is also shown that expanding Minkowski's idea to quantum physics might shed light even on the nature of the quantum object.

Keywords Minkowski spacetime · Worldlines · Worldtubes · Physical laws · Origin of inertia · Four-dimensional stress · Gravity · Quantum object · Discontinuous existence in time

12.1 Introduction

On September 21, 1908 in his talk "Space and Time" Hermann Minkowski proposed a radical change of our views of space and time – "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" [1]. Minkowski arrived at the view that space and time form an independent four-dimensional reality, which he called "the world," by analyzing and successfully revealing the profound meaning of the

V. Petkov (✉)
Science College, Concordia University, Montreal, QC H3G 1M8, Canada
e-mail: vpetkov@alcor.concordia.ca

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_12,

relativity postulate – "the postulate comes to mean that only the four-dimensional world in space and time is given by the phenomena" [1]. He wrote: "the word *relativity-postulate* [...] seems to me very feeble" [1] and preferred "to call it the *postulate of the absolute world*" [1]. This insight also appears to have been helped, as we will see in Section 12.2, by his realization that the consequences of the special theory of relativity are manifestations of the four-dimensionality of the world. In this absolute world, which turned out to be totally counter-intuitive, space and time have equal status – all points of space and all moments of time have equal existence since they form the dimensions of the absolute four-dimensional world of Minkowski, which we now call Minkowski spacetime or simply spacetime.

As the myriad of macroscopic physical bodies which move in the ordinary three-dimensional world are represented by a forever given network of timelike worldlines in spacetime, Minkowski conjectured that there is a close link between physical laws and the network of worldlines: "The whole world seems to resolve itself into such world-lines, and I would fain anticipate myself by saying that in my opinion the laws of physics might find their most perfect expression as interrelations between these world-lines" [1]. This appears to be a natural conjecture since Minkowski seems to have believed[1] that spacetime is not merely an abstract four-dimensional mathematical space, but represents a real four-dimensional world of one temporal and three spatial dimensions. In this world the perceived macroscopic three-dimensional bodies are *real* four-dimensional worldlines or rather worldtubes since we are dealing with spatially extended physical bodies. As physical laws govern the interactions of physical bodies in the three-dimensional world of our perceptions it does appear to follow that physical laws should be directly linked to the network of the bodies' worldtubes in spacetime since only these worldtubes exist there; there are no three-dimensional bodies and no dynamic interactions in the static Minkowski world in the ordinary sense since all interactions are fully realized there.

To my knowledge no consistent attempt has been made so far to examine Minkowski's conjecture. The purpose of this paper is precisely this – to analyze a number of physical laws in order to determine whether or not they are expressions of the interrelations between the worldtubes of macroscopic physical bodies that participate in interactions governed by those physical laws. The analysis will be extended beyond Minkowski's original idea and will try to determine whether quantum laws can be expressed in terms of spacetime structures, not worldlines since quantum objects cannot be represented by worldlines in spacetime. To perform such

[1] Not everyone agrees that Minkowski considered reality to be a four-dimensional world. Although I think he did regard the world as four-dimensional since he appears to have realized that the relativistic effects are manifestations of its four-dimensionality (as will be demonstrated in Section 12.2), it is obvious that what he thought about spacetime does not determine its status of existence.

analyses it is first necessary to address the question of the reality of the macroscopic bodies' worldtubes and ultimately the reality[2] of Minkowski spacetime itself.

Section 12.2 deals with the most important issue raised by Minkowski's paper "Space and Time" – whether the real world is four-dimensional with time being the fourth dimension (which would mean that physical bodies are real timelike four-dimensional worldtubes) and whether the relativistic effects are indeed manifestations of the four-dimensionality of the world as Minkowski anticipated. Section 12.3 demonstrates that the mechanical pre-relativistic and pre-quantum physical laws – Newton's three laws – can be naturally explained as interrelations between worldlines. Section 12.4 explores the internal logic of the concept of gravity as a force (an analysis that appears to have been mostly followed by Einstein) which inescapably leads to the concept of gravity as curvature of spacetime according to which gravitational interaction is expressed as interrelations between geodesic worldlines in curved spacetime. Finally, Section 12.5 examines whether Minkowski's program applies to quantum physics where the quantum objects are not represented by worldlines.

12.2 Minkowski Spacetime and Reality

The major implication of Minkowski's paper "Space and Time" is the issue of the reality of Minkowski spacetime[3]: Is it just a four-dimensional mathematical space or is it representing a real four-dimensional world with one temporal and three spatial dimensions?

To address this issue one should start with an explicit definition of the macroscopic[4] reality (everything that exists at the macro scale). One feature of the world that appears to be self-evident is that it is three-dimensional[5] – macroscopic physical bodies are three-dimensional and space itself is three-dimensional (one can talk

[2] In this paper by "reality of spacetime" I mean the reality of the four-dimensional world as envisioned by Minkowski in which time is entirely given as the fourth dimension and macroscopic physical bodies are four-dimensional worldtubes. The issue of reality of spacetime in terms of the debate absolutism (substantivalism) versus relationism will be mentioned only once in Section 12.4 and will not be discussed here.

[3] It does not matter whether we talk about Minkowski spacetime or any other relativistic spacetime if we ask the question of their reality. What is common to all spacetimes is their four-dimensionality. So by asking whether spacetime is real we ask whether the world is four-dimensional. A real four-dimensional world could be flat (represented by Minkowski spacetime) or curved (represented by another relativistic spacetime).

[4] Relativity describes mostly the macro scale of the world and does not fully apply at the quantum level since its equations of motion manifestly fail to describe the behaviour of quantum objects. The equations of motion of relativity (special or general, depending on the case) govern the behaviour of the perceived macroscopic three-dimensional bodies. That is why it is the macroscopic three-dimensional bodies (not quantum objects) that are represented by timelike worldtubes in relativity.

[5] Even Aristotle regarded reality as a three-dimensional world [2, p. 359].

about distances between bodies in terms of projections only along three mutually orthogonal directions). Another feature which also appears to be self-evident is that everything that exists, exists only at the present moment which constantly changes. The next question is how what we perceive as a three-dimensional space and three-dimensional bodies exist at the present moment. The first, naive view of reality could be defined in the following way – what exists is everything that we *see* (or can in principle see) *simultaneously* at the moment "now"; we "see" space through the distances between objects. So, the four key defining features of the first naive view of reality can be summarized in four words – three-dimensional, see, simultaneously, and "now".

However, when Rømer determined that the speed of light was finite in 1676 it became clear that what we see simultaneously at the moment 'now' is all past since light needs some time to reach our eyes. It turned out that the space we believe we "see" at the present moment does not constitute even a space since it consists of space volumes (around and between bodies) which correspond to different past moments, whereas space is defined in terms of *simultaneity* – as all space points at a given moment of time.

The view of reality established after Rømer's discovery recognized as real everything that *exists simultaneously* at the moment "now". The defining features of this view had not changed significantly and could be summarized again in four words – three-dimensional, exist, simultaneously, and "now". One of these features – the reference to a single moment of time (the present moment) – is an assumption that is based *solely* on the fact that we are aware of ourselves and the world only at that moment. However, it is evident that it does not necessarily follow from this fact that the whole world also exists only at the present moment.[6]

On this view of reality, called *presentism*, all that exists is the present. The present itself is defined as the three-dimensional world – as everything that exists *simultaneously* at the present moment. This pre-relativistic view is still widely accepted not only by the general public but also by some scientists and philosophers. This is a disturbing fact since presentism clearly contradicts the theory of relativity. As the present is defined in terms of *absolute simultaneity* – as everything that exists simultaneously at the moment "now" – if it were only the present that existed, it follows that all observers in relative motion would share the same present (the same set of simultaneous events) which would mean that simultaneity is absolute in contradiction with special relativity.

Then the natural question is: "What is reality according to special relativity?" Or more precisely: "What is the dimensionality of the world at the macroscopic scale where special relativity is fully[7] applicable?" Apparently Minkowski had asked him-

[6] We are aware of ourselves at a specific location in space, but we do not assume that everything that exists, also exists solely at that location.

[7] As indicated in footnote 4 special relativity is a macroscopic theory in its *entirety* and does not fully apply at the quantum level.

self such questions[8] and arrived at the relativistic view of reality – what exists is an absolute four-dimensional world consisting of one temporal and three spatial dimensions. As mentioned in the Introduction he did that by analyzing the physical meaning of the relativity principle – "physical laws would be expressed in exactly the same way by means of x', y, z, t' as by means of x, y, z, t" [1], where x', y, z, t' and x, y, z, t represent the systems of reference of two observers moving uniformly with respect to each other along their x-axes. Minkowski seems to have recognized the far reaching implications of one specific fact following from Einstein's formulation of special relativity – that the times t and t' should be treated equally. The first immediate implication[9] is that *different* times mean *different* classes of simultaneous events (relativity of simultaneity). And as space is defined in terms of *simultaneity* – the class of simultaneous events corresponding to a given moment of time – it follows that the primed observer should "define space by the manifold of the three parameters x', y, z" [1]. Therefore, the two observers in relative motion have not only different times but also *different* three-dimensional spaces.[10] That is why Minkowski's conclusion was unavoidable: "We would then have in the world no longer *the* 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. You see why I said at the outset that space and time are to fade away into shadows, and that only a world in itself will subsist." [1].

---

[8] In my view a clear indication of this is the fact that he discussed the description of what exists – he used the word "the *world*" (in which we can "imagine that everywhere and everywhen there is something perceptible" [1]; that "only a world in itself will subsist" [1]), the word "universe" (which is filled with worldlines of "something perceptible" [1]), "our concept of nature" [1], "four-dimensional physics" [1], and physical objects which are represented by four-dimensional bands [1].

[9] As, on the pre-relativist view, a class of simultaneous events corresponds to each moment of the absolute time, it does follow that two times entail different classes of simultaneous events corresponding to each moment of the two times. In other words, at each moment of their times two observers in relative motion will have different classes of simultaneous events.

[10] Minkowski noticed that "neither Einstein nor Lorentz made any attack on the concept of space" [1]. This is an undeniable triumph of a great mathematician over two great physicists. To realize that different times imply different spaces as well, could have been perhaps also realized by Einstein especially given the fact that Minkowski realized it almost immediately after Einstein insisted that the times t and t' of two observers in relative motion should be treated on equal footing. The omission to notice that different times imply different spaces is more easily explainable in the case of Lorentz. He did not regard the times t and t' of two observers in relative motion as equal since he did not believe t' represented anything real. This meant that an introduction of another space for the primed observer would not be justified. In 1915, in a note added in the second edition of his book "The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat" [8] Lorentz himself described the failure of his attempts to formulate properly the theory of relativity: "The chief cause of my failure was my clinging to the idea that the variable t only can be considered as the true time and that my local time t' must be regarded as no more than an auxiliary mathematical quantity. In Einstein's theory, on the contrary, t' plays the same part as t; if we want to describe phenomena in terms of x', y', z', t' we must work with these variables exactly as we could do with x, y, z, t."

Now the profound meaning of the relativity postulate as revealed by Minkowski becomes completely evident – physical laws are the same in all inertial reference frames in relative motion because each inertial frame has its own space and time and each inertial observer describes the physical laws in his or her space and time exactly like any other inertial observer. Minkowski had realized that many spaces (many classes of simultaneous[11] events) and many times could not exist in a three-dimensional world (i.e., if there existed just *one* space). That would be possible, he argued, only in an absolute four-dimensional world with one temporal and three spatial dimensions, where the spaces of different inertial frames can be regarded as three-dimensional "cross-sections" of it. That is why Minkowski preferred to call the relativity postulate "the *postulate of the absolute world*".

As the principle of relativity in its original formulation given first by Galileo [9] and then generalized by Poincaré [10] and Einstein [11] states that absolute uniform motion cannot be detected, Minkowski's analysis provided the answer to the difficult question that appeared to follow from that impossibility "Why does absolute uniform motion (i.e., uniform motion in the absolute space) not exist?" The answer based on Minkowski's insight turned out to be radical – absolute uniform motion does not exist because no *single* and therefore no *absolute* space exists; what exists is "an infinite number of spaces", which in turn is possible in an absolute four-dimensional spacetime where there is no motion at all (not only absolute uniform motion). What we perceive as motion of three-dimensional bodies is in fact a set of changing with time images of three-dimensional "cross-sections" of the forever given four-dimensional worldtubes of macroscopic physical bodies (since there are no three-dimensional bodies in spacetime).

The conclusion that Minkowski spacetime (sometimes also called a block universe) represents a real four-dimensional world is not merely a possible interpretation of relativity. It is the only interpretation that does not contradict the experimental evidence, which supports relativity.[12] In fact, the support from the experimental evidence is so strong that unequivocally confirms what Minkowski appeared to have realized – that *special relativity is impossible in a three-dimensional*

[11] It appears certain that Minkowski realized that relativity of simultaneity was impossible in a three-dimensional world and regarded it, along with length contraction, as manifestations of the four-dimensionality of the world. So, taken even alone, relativity of simultaneity is sufficient to prove the reality of spacetime.

[12] More precisely, the world at the macroscopic scale should be at least four-dimensional (with one temporal and three spatial dimensions) in order to avoid a direct contradiction with relativity. A model of a growing block universe introduced in 1923 by Broad [13] with some most recent versions [14, 16] has been regarded as an alternative to Minkowski spacetime that does not contradict relativity either. However, by explicitly assuming that the existence of physical bodies is absolute it becomes evident that the growing block universe model also contradicts relativity – the hypersurface (no matter of what shape) on which the birthing of events happens constitutes an objectively privileged hypersurface (existence is absolute!) and therefore an objectively privileged reference frame. Why existence must be regarded as absolute and cannot be relativized is briefly explained in Footnote 14.

world[13] [3]. To see why this is so, assume for a moment that it is not the case and that what exists is indeed the three-dimensional world – the present. Then, as it is the only thing that exists, all observers in relative motion should share the *same* present, i.e., the same set of simultaneous events. Therefore, simultaneity would turn out to be absolute, if the world were three-dimensional. It is then immediately seen that relativity of simultaneity does imply the reality of Minkowski space, if each of two observers in relative motion initially accepts the pre-relativistic presentist view (based on the idea of absolute simultaneity) – as the observers have different sets of simultaneous events it follows that they have different presents, i.e., different three-dimensional worlds. But this is only possible[14] if reality is a four-dimensional world, represented by Minkowski spacetime, which makes it possible for the two observers to regard two different three-dimensional "cross-sections" of it as their presents.

Another way to see why the reality of Minkowski's four-dimensional world cannot be successfully questioned is to ask explicitly whether the *experiments* which confirmed the relativistic effects would be possible if the macroscopic physical bodies involved in these experiments were three-dimensional. A definite answer to such a question directly linking experimental results with the dimensionality of macroscopic bodies and ultimately with the dimensionality of the world can settle the issue of the reality of spacetime once and for all. Analyses of length contraction, time dilation, and the twin paradox clearly demonstrate that these effects would be impossible if the macroscopic physical bodies involved in them were three-dimensional [4, 6].

[13] Some might object that such a claim is wrong since special relativity can be equally formulated in a three-dimensional and a four-dimensional language. They might even point out that its original formulation was in a three-dimensional language. Even if one agrees that the two representations of relativity are equivalent in a sense that they correctly describe the relativistic effects, they are entirely different in terms of the dimensionality of the world. Clearly, the world is either three-dimensional or four-dimensional. Therefore only one of the representations of relativity is correct since only one of them adequately represents the world's dimensionality at the macroscopic scale. Moreover, general relativity cannot be adequately represented in a three-dimensional language. Einstein did formulate special relativity in a three-dimensional language, but when Minkowski asked questions about the physical meaning of the relativity principle and length contraction, for example, it became clear that the relativity principle and the relativistic effects are manifestations of the four-dimensionality of the world.

[14] Strictly speaking, it appears that the three-dimensionalist view (presentism) can be preserved and made compatible with relativity of simultaneity if existence (like motion and simultaneity) is also relativized. In such a case each of two observers in relative motion will acknowledge only the existence of his or her own present and will deny the existence of the other observer's present. However, this possibility is ruled out when relativistic situations not involving relativity of simultaneity are analyzed. For instance, the twin paradox (which is an *absolute* relativistic effect) would be impossible if the twins existed only at their present moments as three-dimensional bodies as required by such a relativized version of presentism [4, 6]. Also, relativization of existence is unquestionably ruled out by taking into account (i) the existence of accelerated observers in special relativity [4] and (ii) conventionality of simultaneity [5].

Consider as an example length contraction [3]. Assume that two observers A and B in relative motion measure the same meter stick which is at rest with respect to A. As a spatially extended three-dimensional body is defined in terms of simultaneity – as "all its parts which exist *simultaneously* at a given moment of time" – relativity of simultaneity implies that the two observers measure *two* different three-dimensional objects: A and B have different sets of simultaneous events, which means that two different three-dimensional meter sticks (two different sets of simultaneously existing parts of the meter stick), one of which is shorter, exist for them. This relativistic fact – while measuring the *same* meter stick two observers in relative motion measure *two* different three-dimensional meter sticks of different lengths – reveals the deep physical meaning of length contraction first realized by Minkowski: length contraction is a manifestation of the four-dimensionality of the meter stick. The meter stick is not the three-dimensional object of our perceptions, but a four-dimensional worldtube, which consists of the ordinary three-dimensional meter stick at *all* moments of its history. The meter stick's worldtube must be a real four-dimensional object in order to allow A and B to regard two different three-dimensional "cross-sections" of it (of different lengths) as their three-dimensional meter sticks. Clearly, what is "the same meter stick" is the meter stick's worldtube.

It should be stressed that if the macroscopic physical objects and the world were three-dimensional, this effect would be *impossible* – if the meter stick's worldtube were not real, this would mean that A and B would measure the *same* three-dimensional meter stick (the same set of simultaneously existing parts of the meter stick); therefore the observers would have a common class of simultaneous events in contradiction with special relativity.

I believe even the concise arguments in this section convincingly demonstrate that Minkowski spacetime represents a real four-dimensional world in which the entire histories of all *macroscopic* three-dimensional bodies of our perceptions are realized as the bodies' worldtubes. What we perceive as interactions between three-dimensional bodies which are governed by physical laws are in reality a forever given network of worldtubes. Therefore Minkowski's program does appear to be a natural conjecture since if we can identify physical laws by studying the apparent interactions of the observed three-dimensional reflections of physical bodies, we should be able to recognize those laws by examining the network of worldtubes and to realize that the physical laws are merely interrelations between worldtubes.

Minkowski himself demonstrated how his program works in electrodynamics (in the cases of "the elementary laws formulated by A. Liénard and E. Wiechert" and "the ponderomotive action of a moving point charge on another moving point charge[15]" [1]) and in relativity by revealing the profound physical meaning of length

[15] Minkowski did not hide his satisfaction at the application of his program [1]:

> "When we compare this statement with previous formulations of the same elementary law of the ponderomotive action of a moving point charges on one another, we are compelled to admit that it is only in four dimensions that the relations here taken under consideration

contraction[16] – the relativistic length contraction of physical bodies and distances is a manifestation of the four-dimensionality of the bodies and the world. In the spacetime diagram depicted in Fig. 1 of his paper [1] he represented two Lorentzian electrons in relative motion by two four-dimensional bands in spacetime which form an angle between them. Each of these electrons can be regarded at rest when a separate time axis is introduce along the band of each electron. Then the spatial axis of each electron intersects the two bands in two "cross-sections" one of which represents the proper length of the electron at rest and the other – the contracted length of the moving electron. Minkowski demonstrated that the two "cross-sections" are related by the correct formula of relativistic length contraction and concluded: "But this is the meaning of Lorentz's hypothesis of the contraction of electrons in motion" [1].

It turns out that Minkowski's program is the only tool for revealing the deep physical meaning of the relativistic effects. For example, the twin paradox (as an absolute relativistic effect) was fully understood and explained only when it was realized that it is the triangle inequality in spacetime; if it is assumed that the worldtubes of the twins (which form an idealized triangle in spacetime) were not real four-dimensional objects and that the twins existed only at their present moments as the ordinary three-dimensional bodies of our perceptions, this relativistic effect would be impossible ([6], Ch. 5). By following the same approach and assuming that the worldtubes of two clocks in relative motion were not real and that the clocks existed only at their present moments as the ordinary three-dimensional clocks, it can be shown that another relativistic effect – time dilation – could not exist either ([6], Ch. 5).

> reveal their inner being in full simplicity, and that on a three dimensional space forced upon us from the very beginning they cast only a very tangled projection".

Another example of how Minkowski's program may provide a simpler and more adequate picture of some controversial issues in electrodynamics is the debate on whether or not a uniformly accelerating charge radiates (see, for example, [17], Ch. 17 and the references therein). An arbitrarily accelerated charge radiates, whereas a charge moving with constant velocity, i.e., by inertia, does not. In terms of worldlines, a charge whose worldline is curved (deformed) radiates, whereas a charge whose worldline is straight (undeformed) does not. As the worldline of an uniformly accelerating charge is also deformed the qualitative answer to the question whether or not such a charge should radiate is obviously affirmative. Then a quantitative answer is expected to be also affirmative ([17], Ch. 17 and the corresponding references therein).

[16] Minkowski specifically pointed out that length contraction is not "a consequence of resistance in the ether" [1]:

> According to Lorentz any moving body must have undergone a contraction in the direction of its motion. In particular, if the body has the velocity v, the contraction will be of the ratio
>
> $$1 : \sqrt{1-\frac{v^2}{c^2}}\,.$$
>
> This hypothesis sounds extremely fantastical. For the contraction is not to be thought of as a consequence of resistance in the ether, but simply as a gift from above, as an accompanying circumstance of the circumstance of motion.

12.3 Newton's Three Laws and Worldlines

This section deals with Newton's three laws of motion and some old difficulties involved in our understanding of uniform and accelerated motion that are still without a satisfactory explanation. We will see that Newton's laws are in fact basic statements about worldtubes in Minkowski spacetime which provide a natural resolution of those difficulties.

The three laws of motion in Newton's original formulation are [18]:

1. Every body perseveres in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed thereon.
2. The alteration of motion is ever proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed.
3. To every action there is always opposed an equal reaction; or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.

The main difficulties in understanding inertial (uniform) and accelerated motion can be summarized as:

- Why can the state of rest and of uniform motion along a straight line (i.e., motion with constant velocity) not be experimentally distinguished? In other words, why can motion with constant velocity not be discovered?
- Why does a body moving with constant velocity appear to move on its own and can move forever if nothing prevents it from doing so?
- Why does a body resist any change in its state of motion with constant velocity? In other words, why can accelerated motion be detected?

A physical body is a forever given four-dimensional timelike worldtube in spacetime. A straight worldtube represents a body which moves with a constant velocity with respect to an inertial reference frame. We can choose the time axis of another inertial reference frame along the worldtube of this body, which means that the body will appear at rest in the new frame.[17] This provides an answer to the first question above – the worldtubes of a body at rest in an inertial reference frame and of another body moving with constant velocity with respect to the first body

---

[17] Minkowski dealt with this situation in the following way [1]:

> We now want to introduce this fundamental axiom:
>
> *The substance existing at any world-point may always, with the appropriate fixation of space and time, be looked upon as at rest.*
>
> The axiom signifies that at every world-point the expression
>
> $$c^2dt^2 - dx^2 - dy^2 - dz^2$$
>
> always has a positive value, or, what comes to the same, that any velocity v always proves less than c. Accordingly c would stand as the upper limit for all substantial velocities, and that is precisely what would reveal the deeper significance of the magnitude c.

are both straight (undeformed). The fact that all straight timelike worldtubes in spacetime are equivalent explains why the states of rest and motion with constant velocity cannot be distinguished – one can choose the time axis of an inertial reference frame along a given (arbitrarily chosen) straight timelike worldtube, which means that the body represented by that worldtube will appear at rest, whereas the bodies represented by other straight timelike worldtubes, forming an angle with the first worldtube, will appear in motion with constant velocity. Choosing the time axis along another straight timelike worldtube will provide a different description in three-dimensional language of the set of worldtubes according to which the first body will appear moving with constant velocity with respect to the new inertial frame.

When described in three-dimensional language a body, whose worldtube is straight, appears to *move on its own* with constant velocity relative to a an inertial reference frame. Since Galileo [9] (who disproved Aristotle's view that "everything that is in motion must be moved by something" [19], p. 167) it has been taken as a postulate (supported by the experimental evidence) that a body moving with constant velocity appears to move on its own which could continue forever if the body did not encounter any obstacles. This motion by inertia, which Newton postulated as his first law of motion, has never been given an explanation. The explanation became possible only after Minkowski's four-dimensional representation of special relativity, but it is completely unexpected – a body moving by inertia does not move on its own; in reality, it does not move at all since the body is a four-dimensional worldtube in the static Minkowski world. A body moving by inertia only *appears* to move on its own (and to be in motion), but it is only due to our perception reflecting the fact that the body's worldtube is straight – at each moment we perceive the light reflection from a different three-dimensional "cross-section" of the inclined straight worldtube of the body we observe and it *appears* that a three-dimensional body moves uniformly. So Newton's first law turns out to be a statement of the existence of straight timelike worldtubes in spacetime.

Whenever we talk about inertia we often mean two aspects: (i) a body moves with constant velocity *on its own* (Newton's first law), and (ii) a body *resists* any change in its state of motion with constant velocity (Newton's second law). As we saw the first aspect of inertia reflects the fact that there exist straight timelike worldtubes in spacetime. The second aspect of inertia as well as Newton's second and third laws and the last question in the list above can be best explained if we try to understand why an accelerating body resists its acceleration. For centuries all attempts to explain that resistance had been mostly seeking its origin in space, but there had been an obvious problem – if space is the cause of that resistance, then why does the motion of a body moving by inertia not be resisted by space?

Minkowski's program offers an amazing explanation of the resistance every accelerating body encounters. First, note that the worldtube of an accelerating body is *deformed* (not straight). Second, after the realization that a macroscopic physical body is a *real* four-dimensional timelike worldtube in spacetime it is natural to assume that the deformed worldtube of an accelerating body (like a deformed three-dimensional rod) *resists its deformation*. The four-dimensional stress arising in the deformed worldtube of an accelerating body gives rise to a static restoring force which tries to restore the worldtube to its initial undistorted shape. This restoring force manifests itself as the inertial force to which the accelerating body is subjected.[18] Therefore, the resistance an accelerating body encounters does not come from space but originates in the body *itself*, more precisely in the body's deformed worldtube.

Hence the answer to the question "Why does a particle resist any change in its state of motion with constant velocity?" turns out to be quite unexpected. I guess Minkowski would have been thrilled to realize that the very existence of inertia (especially its second aspect) is another manifestation of the reality of worldtubes and spacetime itself.

Now the answers to the questions of why an accelerated motion can be detected experimentally, whereas motion by inertia cannot, are also clear. Accelerated[19] motion can be detected since the worldtube of an accelerated body is deformed and resists its deformation. This four-dimensional resistance is manifested as the resistance the body offers to its acceleration. It is through this resistance any accelerated motion can be experimentally detected. The worldtube of a body moving with constant velocity is *not* deformed, which means that the body offers no resistance to its inertial motion. That is why inertial (i.e., non-resistant) motion cannot be discovered.

Newton's second law turns out to be a statement that in order to have a statically curved worldtube (or, in three-dimensional language, to accelerate a body) a force should be applied (to overcome the worldtube's static resistance to deformation). Such a force comes from another worldtube that statically curved (deformed) the first one. As the two worldtubes are mutually curved by each other in the static Minkowski spacetime, each of them resists its deformation caused by the other worldtube. Therefore, what is an external force for one worldtube (for the accelerated body) is a resistance force for the other worldtube (an inertial force acting back on the body that accelerates the first one) and vice versa. As a result of this symmetry the two forces have equal magnitudes and opposite directions, which is Newton's third law.

---

[18] Calculations of the static restoring force arising in the deformed worldtube of an accelerating body show that it does have the form of the inertial force ([6], Ch. 10).

[19] There still exists some confusion on whether acceleration is absolute in relativity. When Minkowski's program is employed it becomes clear that absolute acceleration does not imply absolute space (relative to which a body accelerates). The acceleration of a body is absolute in a sense that one does not need another body *relative* to which the first one accelerates; the acceleration of the first body can be determined in an *absolute* way by detecting the resistance the body offers to its acceleration. In special relativity any acceleration is absolute – a body whose worldtube is deformed accelerates (in three-dimensional language) in an absolute fashion. In general relativity there exist two types of acceleration: (i) absolute acceleration when a body resists its acceleration (the body's worldtube is deformed), and (ii) relative acceleration caused by the so called geodesic deviation (the bodies subjected to relative acceleration do not resist their acceleration since their worldtubes are geodesic, i.e., curved, but not deformed) ([6], Sec. 8.1).

So, Minkowski's program works perfectly in the case of Newtonian physics as well – Newton's three laws are merely statements about the existence of straight timelike worldtubes in Minkowski spacetime which, like ordinary tubes or rods, resist their *static* deformation.

12.4 Spacetime and Gravity

In this section we will see how Minkowski's program could have been consistently applied in order to arrive at the general theory of relativity. Let us start with the question: Could Minkowski have discovered general relativity by following his own program if he had lived longer? This seems unlikely if he had followed his approach to achieving new results in theoretical physics as revealed in his talk "Space and Time". There Minkowski implied that special relativity could have been discovered on the basis of mathematical considerations alone, but made it clear that that possibility was realized only after the theory was discovered [1]:

> Such a premonition would have been an extraordinary triumph for pure mathematics. Well, mathematics, though it now can display only staircase-wit, has the satisfaction of being wise after the event, and is able, thanks to its happy antecedents, with its senses sharpened by an unhampered outlook to far horizons, to grasp forthwith the far-reaching consequences of such a metamorphosis of our concept of nature.[20]

Clearly, as a mathematician Minkowski believed that pure mathematics could play a leading role in discoveries in physics. However, this view is not completely shared even by other mathematicians interested and involved in theoretical physics. Here is what Hermann Weyl wrote on this issue [20]:

> All beginnings are obscure. Inasmuch as the mathematician operates with his conceptions along strict and formal lines, he, above all, must be reminded from time to time that the origins of things lie in greater depths than those to which his methods enable him to descend. Beyond the knowledge gained from the individual sciences, there remains the task of *comprehending*. In spite of the fact that the views of philosophy sway from one system to another, we cannot dispense with it unless we are to convert knowledge into a meaningless chaos.

It is true that mathematical considerations helped Minkowski to realize that space and time are different dimensions of an absolute underlying reality – spacetime. But physical considerations played a crucial role. Minkowski himself admitted that "The views on space and time which I wish to lay before you have sprung from the soil of experimental physics" [1].

I think purely mathematical approach could not have succeeded in discovering that gravity is not a force, but a manifestation of the curvature of spacetime.

[20] Here we again see another indication that Minkowski does not seem to have regarded the four-dimensional world uniting space and time as a mathematical space – he talks about "such a metamorphosis of our concept of *nature*" (italics added).

Minkowski's treatment of gravity at the end of his paper "Space and Time" is not very promising since he still regarded the gravitational attraction as a force [1]:

> In mechanics as reformed in accordance with the world-postulate, the disturbing lack of harmony between Newtonian mechanics and modern electrodynamics disappears of its own accord. Before concluding I want to touch upon the attitude of Newton's law of attraction toward this postulate. I shall assume that when two mass points m, m_1 follow their worldlines, a motive force vector is exerted by m on m_1, of exactly the same form as that just given for the case of electrons, except that $+mm_1$ must now take the place of $-ee_1$.

I will now outline a possible conceptual analysis of physical facts of gravitational physics (known in 1908), which applies Minkowski's program. Such an analysis, which, as Minkowski put it, can now display only staircase-wit, demonstrates how naturally and smoothly one arrives at general relativity when the implications of Minkowski's idea of the absolute four-dimensional world are analyzed and the issue of the reality of worldtubes is taken seriously.

A conceptual analysis of Newton's gravitational theory could and should have revealed, long before Einstein realized it, that there are problems with Newton's notion of gravitational force. The first of those problems was realized by Einstein most probably in November 1907 and this insight set him on the path toward his theory of general relativity (quoted from [21]):

> I was sitting in a chair in the patent office at Bern when all of a sudden a thought occurred to me: "If a person falls freely he will not feel his own weight." I was startled. This simple thought made a deep impression on me. It impelled me toward a theory of gravitation.

Einstein was so impressed by this insight that he called it the "happiest thought" of his life [21].

And indeed if the fall of a body in a gravitational field is conceptually analyzed it becomes clear that there is a problem with Newton's explanation that the body is falling because it is subjected to a gravitational force. According to Newton's second law $\mathbf{F} = m\mathbf{a}$ a force is necessary to accelerate a body since the body resists its acceleration (and the force should overcome that resistance). Therefore, a falling body should be subjected to a force $\mathbf{F}_g = m\mathbf{g}$, which Newton called a gravitational force, since it forces the body to accelerate with an acceleration $\mathbf{g}$.

However, the gravitational force is very different from the contact forces described by Newton's second law since it is an "action at a distance" force. Newton himself appeared to have had a lot of difficulty understanding the nature of such a non-contact force. In a letter to Richard Bentley Newton wrote [22]:

> It is inconceivable, that inanimate brute matter, should, without the mediation of something else, which is not material, operate upon and affect other matter without mutual contact, as it must be, if gravitation, in the sense of Epicurus, be essential and inherent in it. And this is one reason why I desired you would not ascribe innate gravity to me. That gravity should be innate, inherent, and essential to matter, so that one body may act upon another at a distance through a vacuum, without the mediation of any thing else, by and through which their action and force may be conveyed from one to another, is to me so great an absurdity, that I believe no man, who has in philosophical matters a competent faculty of thinking, can ever fall into it. Gravity must be caused by an agent acting constantly according to certain laws; but whether this agent be material or immaterial, I have left to the consideration of my readers."

What is also strange with this force is what Galileo discovered through conceptual analysis[21] (and possibly through experiment[22] as well) of Aristotle's claim that heavy bodies fall faster than light ones. Galileo concluded that all bodies, no matter heavy or light, fall at the same rate, or in Newton's terms, with the *same* acceleration. This experimental fact is so bizarre that it needed another Galileo – another master of thorough conceptual analyses – to reveal what that fact has been trying to tell us. It is unclear whether Einstein, whose thought experiments made him a second Galileo in this respect, had extracted some valuable information from that fact.

A thorough analysis of Galileo's discovery could have revealed an interesting similarity between that discovery and inertial motion. According to Newton's first law of motion (and Galileo's own experiments [9] which had led him to the idea of inertial motion) different bodies move with the *same* velocity *by inertia* no matter whether they are heavy or light. So if heavy and light bodies fall with the same acceleration it is tempting the say that they move by inertia[23] and because of this it does not matter whether they are heavy or light. However, the problem is obvious – how could they move by inertia if they accelerate?

But if one does not give up and continues to analyze Galileo's strange discovery and even performs other simple experiments an insight will almost certainly enlighten such a person. When we allow drops of water to fall we can see that they are not deformed which demonstrates that the drops do not resist their downward

[21] Through Salviati Galileo demonstrated to Simplicio that the assumption – a heavy body falls faster than a light body – leads to a contradiction ([23], p. 446):

> If then we take two bodies whose natural speeds are different, it is clear that on uniting the two, the more rapid one will be partly retarded by the slower, and the slower will be somewhat hastened by the swifter... But if this is true, and if a large stone moves with a speed of, say, eight while a smaller moves with a speed of four, then when they are united, the system will move with a speed less than eight; but the two stones when tied together make a stone larger than that which before moved with a speed of eight. Hence the heavier body moves with less speed than the lighter; an effect which is contrary to your supposition. Thus you see how, from your assumption that the heavier body moves more rapidly than the lighter one, I infer that the heavier body moves more slowly.

[22] Again through Salviati Galileo implies that he did perform experiments with heavy and light bodies ([23], pp. 447–448):

> Aristotle says that "an iron ball of one hundred pounds falling from a height of one hundred cubits reaches the ground before a one-pound ball has fallen a single cubit." I say that they arrive at the same time. You find, *on making the experiment* [italics added], that the larger outstrips the smaller by two finger-breadths, that is, when the larger has reached the ground, the other is short of it by two finger-breadths; now you would not hide behind these two fingers the ninety-nine cubits of Aristotle, nor would you mention my small error and at the same time pass over in silence his very large one.

[23] Asking the question of whether a falling body moves non-resistantly, on its own, does not appear to be intellectually unachievable even because Aristotle had already regarded the motion of falling bodies natural or not forced ([19], pp. 136, 199).

acceleration.[24] We can test that amazing observation by jumping from a hight, for example, and will also observe that our bodies do not resist our fall. Also, if a balloon filled with water is attached to a string, so it is prevented from falling, its shape deforms, which demonstrates that the balloon resists its state of rest in the Earth's gravitational field. However, if the balloon is allowed to fall, its shape becomes spherical, which is a clear indication that the falling balloon does not resist its fall.

These observations imply that a falling body offers no resistance to its acceleration. Therefore, it is not unthinkable to imagine that one can arrive through analyzing gravitational phenomena (not necessarily through insight like Einstein) at the conclusion that a falling body is not subjected to any gravitational force, which would otherwise be necessary if the body resisted its fall. This would mean that the falling body moves non-resistantly, by inertia. But again – how could that be since it accelerates?

Taking seriously the existence of worldtubes and following Minkowski's program makes it possible to decode the message hidden in Galileo's discovery that both heavy and light bodies fall with the same acceleration. This message turns out to be profound – gravity is a manifestation of the curvature of spacetime.

And indeed the new view of gravity follows unavoidably (not just naturally) – the worldtube of a falling body should be curved (to reflect the fact that the body accelerates), but not deformed (to account for the fact that the body does not resist its acceleration). Such a worldtube cannot exist in the flat Minkowski spacetime since if a worldtube is curved there it is also deformed. Only in curved spacetime the worldtubes of bodies moving by inertia (non-resistantly), called geodesics, are *curved but not deformed*. So, heavy and light bodies fall with the same acceleration since they indeed move by inertia; their acceleration is a manifestation of the curvature of spacetime since their worldtubes are both curved and not deformed.

In fact, a second problem with the Newtonian notion of gravitational force could have been realized before the advent of general relativity, if this notion had been rigorously examined. According to Newton's gravitational theory a body supported in a gravitational field (say, placed on a table) is subjected to the gravitational force $\mathbf{F}_g = m\mathbf{g}$, where $\mathbf{g}$ is the acceleration due to gravity. The obvious fact here is that the body does not accelerate and does not even move. But how could then a force act on the body if it does not accelerate?

---

[24] Galileo (through Salviati) virtually arrived at the conclusion that a falling body does not resist its fall ([23], p. 447):

> But if you tie the hemp to the stone and allow them to fall freely from some height, do you believe that the hemp will press down upon the stone and thus accelerate its motion or do you think the motion will be retarded by a partial upward pressure? One always feels the pressure upon his shoulders when he prevents the motion of a load resting upon him; but if one descends just as rapidly as the load would fall how can it gravitate or press upon him? Do you not see that this would be the same as trying to strike a man with a lance when he is running away from you with a speed which is equal to, or even greater, than that with which you are following him? You must therefore conclude that, during free and natural fall, the small stone does not press upon the larger and consequently does not increase its weight as it does when at rest.

It is ironic that even today this question still can confuse some physics students. One may hear "explanations" such as this one – the body does not accelerate (and does not move) since there are two exactly counterbalancing forces at work: gravity pulling the body down and the table pushing it up. So the net force acting on the body is zero, which means that the body should not accelerate (and not move).

The error here, I think, is obvious – this "explanation" answers a wrong question: "Why does the body not accelerate and not move?" The question we asked is: "Why is there a gravitational force $\mathbf{F}_g = m\mathbf{g}$ (that is balanced by the normal reaction force coming from the table), if the body does not accelerate?" The question is about the very existence (origin) of the gravitational force (that causes the normal reaction force).

Had the two problems with the notion of gravitational force – that the Newtonian gravitational theory does not have an answer to the above question and that a falling body does not offer any resistance to its acceleration – been realized before 1908 and Minkowski's program had been taken seriously, the discovery of general relativity could have been a natural and unavoidable discovery. Now many physicists feel that general relativity appeared so quickly after special relativity (which made Einstein ask what is the speed of propagation of gravity) mostly due to the lucky fact that we had Einstein.

Now general relativity regards gravity as a manifestation of the curvature of spacetime[25] and provides a consistent no-force explanation of gravitational interaction of bodies which follow geodesic paths, i.e., which are represented by geodesic worldtubes. But it is silent on the nature of the very force that has been regarded as gravitational – the force acting upon a body at rest in a gravitational field.

However, when the reality of worldtubes is taken into account and Minkowski's program is employed the picture provided by general relativity becomes complete and fully consistent. The worldtube of a body falling in a gravitational field is geodesic and the body does not resist its fall since its worldtube is not deformed.[26]

[25] This will be the only place in the paper where the debate over the ontological status of spacetime itself (i.e., the debate substantivalism versus relationalism) will be briefly mentioned. Spacetime must exist in order to explain gravity – if spacetime were a non-entity, no matter how glorious, it could not curve; what does not exist does not possess real properties such as curvature, which manifest itself in the real gravitational interaction.

[26] If we consider a body falling toward the surface of the Earth, according to general relativity it is the surface that accelerates since its worldtube is deformed (not geodesic), whereas the falling body moves by inertia (non-resistantly) and its worldtube is geodesic (not deformed). One can see from here why the equivalence principle works. If a body falls toward the floor of an accelerating rocket, it is the floor that accelerates (its worldtube is deformed), whereas the falling body moves by inertia (its worldtube is geodesic). However, if a body falls toward the center of the Earth, both the worldtube of the body and the worldline of the Earth's center are geodesic and no true acceleration (deformation of a worldline) is involved. The apparent acceleration between the body and the Earth's center is caused by the fact that there are no straight and no parallel worldlines in curved spacetime. The body and the Earth's center only appear to accelerate relative to each other; the rate of change of the distance between them is given by the equation of geodesic deviation [24].

But when the body is prevented from falling its worldtube is deviated from its geodesic shape and therefore deformed.

The restoring force that arises in the body's deformed worldtube turns out to be inertial since it has the same origin as in the case of an accelerating body. But in this case it is traditionally called the gravitational force. Therefore regarding the worldtubes of physical bodies as real four-dimensional objects demonstrates that the force acting on a body supported in a gravitational field is indeed *inertial* [25], which naturally explains why "there is no such thing as the force of gravity" in general relativity [26] and why inertial and gravitational forces (and masses) are equivalent ([6], Ch. 10).

12.5 Spacetime and Quantum Physics

At first sight it appears that Minkowski's program cannot be applied in quantum physics for two reasons. First, we do not know what the quantum object (e.g., an electron or a photon) is, and according to the standard interpretation of quantum mechanics we cannot say or even ask anything about the quantum object between measurements. In this sense, I think, Einstein was right that quantum mechanics is essentially incomplete. However, it is unrealistic to assume that an electron, for example, does not exist between measurements. But if it exists, it is something and we should know what that something is.

Second, although it is not clear what an electron is, it is certain what the electron is *not* – it is not a worldline in spacetime. Then, how could Minkowski's program – physical laws might find their most perfect expression as interrelations between worldlines – be applied? Here we will go beyond this program and ask whether physical laws can be expressed in terms of some spacetime structures, not worldlines.

Two things appear unquestionable: (i) relativity does not fully apply at the quantum level since its equations of motion do not describe adequately the behaviour of quantum objects, and (ii) spacetime has the same status in the quantum world as in the macroscopic world – it is the underlying reality at both levels. Therefore quantum physics should provide a spacetime model of the electron and of all quantum objects.

In an attempt to get an insight into what the spacetime model of an electron might be, let us first see why an electron is not a worldline. This, for example, can be demonstrated in the case of interference experiments performed with single electrons [27]. In such double-slit experiments accumulation of successive single electron hits on the screen builds up the interference pattern that demonstrates the wave behaviour of *single* electrons. If we look at the screen, we see that every single electron is detected as a localized entity and we are tempted to think that the electron was such an entity before it hit the screen. Our intuition leads us to assume that if the electron hits the screen as a localized entity, it is such an entity at *every* moment of time, which means that the electron exists *continuously* in time as a localized entity. But if this were the case, every single electron would behave as an ordinary particle and should go only through one slit and no interference pattern would be observed on the screen. Therefore, the inescapable conclusion is that *the electron is not a localized entity at all moments of time*, i.e., it is not a worldline in spacetime. So then, what is the electron in spacetime?

The apparent paradox – every single electron must go through both slits (in order to hit the screen where the "bright" fringes of the interference pattern form) but is always detected on the screen as a localized entity – is obviously trying to tell us something about what the electron is. How can this message be decoded? Let us start with the unquestionable facts. Every time an electron is detected it is localized *in space*. The other fact is that an electron, when not measured, is not a worldline. That is, it is not localized in space at all moments of time. What are the alternatives then?

Our intuition might suggest an obvious alternative – when not measured an electron is some kind of a fluid and for this reason it is not localized in space. However, the difficulties with this model are enormous. It is sufficient to mention just two. First, it is unexplainable that an arbitrary fraction of this fluid, i.e. a fraction of the electron charge, has never been measured. Second, when measured the electron fluid must instantaneously collapse into the small location where the electron is detected, which leads to a contradiction with relativity since a physical fluid must move at infinite speed (and also, what is infinite in one frame of reference is not infinite in another).

Is there any other interpretation of the fact that an electron, when not observed, is not localized in space at all moments of time? Or, in other words, how can the fact that an unobserved electron is not a worldline be interpreted? When it is explicitly taken into account that a worldline represents an object that *continuously* exists in time, a possible interpretation becomes almost obvious – if an electron cannot be represented by a worldline, it may mean that it does not exist continuously at all moments of time. This interpretation provides an amazingly symmetric spacetime model of the electron – no matter whether or not an electron is observed, it is always localized both in space and time or, more precisely, it is localized in spacetime. But since it exists discontinuously in time (only at some, not at *all* moments of time) it is not a point-like entity localized just in a single spacetime point.

Perhaps the best way to envisage an electron which does not exist continuously in time is to imagine that its worldline is disintegrated into its constituent four-dimensional points. Such a spacetime model of the electron represents it not as a worldline, but as a class of point-like entities (with non-zero dimensions) scattered all over the spacetime region where the electron wavefunction is different from zero.

In our three-dimensional language such an electron will appear and disappear at a given point in space and appear and disappear at a distant location in space and so on. This does not imply motion faster than light since the electron does not move as an entity which continuously exists in time from one to the another space point. Also, such an electron possesses an internal frequency of appearance and disappearance which could explain the physical meaning of the Compton frequency of the electron. In terms of the spacetime model, the Compton frequency implies that for one second an electron will be represented by 10^{20} point-like entities. When an electron is not measured it is *actually* everywhere in the spacetime region where its wavefunction is different from zero, because its constituents are scattered all over that region. So it becomes clear how such an electron can go through both slits in the double slit experiment. When the first four-dimensional point of an electron falls in a detector it is trapped there due to a jump of the boundary conditions and all its consecutive points also appear in the detector, which means that an electron is always measured as a localized entity.

An important feature of this spacetime model of the quantum object is that the probabilistic behaviour of quantum objects does not contradict at all the relativistic forever given spacetime picture of the world – the *probabilistic distribution* of the four-dimensional points of an electron in the spacetime region where the electron wavefunction is different from zero is *forever given* in spacetime.

The attempt to extend Minkowski's program to the quantum world leads to a spacetime model of the quantum object which allows to view the quantum laws governing the probabilistic behaviour of quantum objects as reflecting the spacetime probabilistic distributions of the constituents of each quantum object.

In these desperate times in quantum physics it is worth searching for a spacetime model of the quantum object, which might provide answers to the difficult and controversial questions in quantum mechanics. In this section we briefly demonstrated how such a model, based on the idea [28] that the quantum objects do not exist continuously in time, can provide a completely different and paradox-free view of quantum phenomena.[27]

12.6 Conclusions

A hundred years after Minkowski presented his paper "Space and Time" we still owe him answers to some deep questions and ideas he outlined in his paper. One of those ideas was that physical laws might find their most perfect expression as interrelations between the worldtubes of physical bodies. He himself demonstrated how this program worked in the case of several examples one of which dealt with the physical meaning of length contraction – Minkowski anticipated that this effect is a manifestation of the reality of the worldbands representing two Lorentzian electrons subjected to reciprocal length contraction.

The purpose of this paper was to examine further whether Minkowski's program can be applied to different areas of physics. It was indeed demonstrated that, when the reality of worldtubes is taken into account, not only can physical laws be regarded as interrelations between worldtubes of macroscopic bodies,

[27] For a more detailed conceptual account of the idea that quantum objects may exist discontinuously in time see ([6] Ch. 6).

but a deeper understanding of the corresponding phenomena can be achieved. It was shown in Section 12.3 that the mechanical pre-relativistic and pre-quantum physical laws – Newton's three laws – can be explained as statements about the existence of straight worldtubes in flat spacetime and interrelations between them; as a bonus Minkowski's program shed some light on the possible origin of inertia. In Section 12.4 the internal logic of the concept of gravity as a force was explored and it was demonstrated that Minkowski's program inescapably leads to the concept of gravity as curvature of spacetime according to which gravitational interaction is expressed as relations between geodesic worldtubes of macroscopic bodies in curved spacetime. Finally, Section 12.5 attempted to expand Minkowski's program to quantum physics and it was shown that it has the potential to shed light even on what the quantum object might be.

References

1. Minkowski, H.: Space and Time. In this volume
2. Aristotle: On the Heavens, Book I, Ch. 1. In: Adler, M.J. (ed.) Great books of the Western world, Vol. 7, pp. 357–405. Encyclopedia Britannica, Chicago (1993)
3. Petkov, V.: On the reality of Minkowski space. Found. Phys. **37** (10), 1499–1502 (2007)
4. Petkov, V.: Relativity, dimensionality, and existence. In: Petkov, V. (ed.) Relativity and the dimensionality of the world. Springer, Berlin, Heidelberg, New York, pp. 115–135 (2007)
5. Petkov, V.: Conventionality of simultaneity and reality. In: Dieks, D. (ed.) The ontology of spacetime. Philos. Found. Phys. Series **4**, 174–185. Elsevier, Amsterdam (2006)
6. Petkov, V.: Relativity and the Nature of Spacetime. Springer, Berlin Heidelberg New York (2005)
7. Petkov, V.: Is there an alternative to the block universe view? In: Dieks, D. (ed.) The Ontology of Spacetime. Philos. Found. Phys. Series, **1**, 207–228. Elsevier, Amsterdam (2006)
8. Lorentz, H.A.: The Theory of Electrons and Its Applications to the Phenomena of Light and Radiant Heat, 2nd edn, p. 321. Dover, Mineola, New York (2003)
9. Galileo, G.: Dialogue Concerning the Two Chief World Systems – Ptolemaic and Copernican, 2nd edn, Chap. 2 (The Second Day). University of California Press, Berkeley (1967)
10. Poincaré, H.: Sur la dynamique de l'électron. Comptes Rendues **140**, (1905) p. 1504
11. Einstein, A.: On the electrodynamics of moving bodies. In: Lorentz, H.A., Einstein, A., Minkowski, H., Weyl, H. (eds.) The principle of relativity: A collection of original memoirs on the special and general theory of relativity, pp. 37–65. Dover, New York (1952)
12. Lorentz, H.A., Einstein, A., Minkowski, H., Weyl, H.: The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity. pp 75–91. Dover, New York (1952)
13. Broad, C.D.: Scientific Thought Routledge and Kegan Paul, London (1923). In: van Inwagen, P., Zimmerman, D. (eds.) Broads defense of the growing block model is reprinted as Ch. 8. in *Metaphysics: The Big Questions*. Blackwell, Malden (1998)
14. Ellis, G.F.R.: Physics in the real universe: Time and spacetime. Gen. Rel. Grav. **38** (2006) 1797–1824, Arxiv:gr-qc/0605049
15. Christian, J.: Absolute being versus relative becoming. In: Petkov, V. (ed.) Relativity and the dimensionality of the World. Springer, Berlin (2007), Arxiv:gr-qc/0610049
16. Sorkin, R.D.: Relativity theory does not imply that the future already exists. In: Petkov, V. (ed.). Relativity and the dimensionality of the World. Springer, Berlin (2007), Arxiv:gr-qc/0703098
17. Lyle, S.N.: Uniformly Accelerating Charged Particles. Springer, Berlin, Heidelberg, New York (2008)

18. Newton, I.: The principia: mathematical principles of natural philosophy. In: Hawking, S. (ed.) On the shoulders of giants, pp. 733–734. Running Press, Philadelphia (2002)
19. Aristotle, Physics Oxford University Press, New York (2008)
20. Weyl, H.: Space-Time-Matter, p. 10. Dover, New York (1952)
21. Pais, A.: Subtle is the Lord: The Science and the Life of Albert Einstein, p. 179. Oxford, Oxford University Press (2005)
22. Bentley, R.: "Letter of Newton to Bentley, Trinity College, Jan. 17, 1692-3." In: *Works of Richard Bentley*, vol. 3 (London 1838) pp. 210–211. Quoted in: Sir Isaac Newton's Mathematical Principles of Natural Philosophy and His System of the World, p. 634. Kessinger Publishing (2003)
23. Galileo Dialogues concerning two sciences. In: Hawking, S. (ed.) On the shoulders of giants, pp. 399–626. Running Press, Philadelphia (2002)
24. Ohanian, H., Ruffini, R.: Gravitation and Spacetime, 2nd edn, p. 343. W.W. Norton, New York, London (1994)
25. Rindler, W.: Essential relativity, 2nd edn, p. 244. Springer, Berlin (1977)
26. Synge, J.L.: Relativity: The General Theory, p. 109. Nord-Holand, Amsterdam (1960)
27. Tonomura, A., Endo, J., Matsuda, T., Kawasaki, T., Exawa, H.: Demonstration of single-electron buildup of an interference pattern. Am. J. Phy., **57**, 117–120 (1989)
28. Anastassov, A.H.: Self-contained phase-space formulation of quantum mechanics as statistics of virtual particles "Annuaire de l'Universite de Sofia St. Kliment Ohridski", Faculte de Physique **81**, 135–163 (1989)

Chapter 13
Time as an Illusion

Paul S. Wesson

Abstract We review the idea, due to Einstein, Eddington, Hoyle and Ballard, that time is a subjective label, whose primary purpose is to order events, perhaps in a higher-dimensional universe. In this approach, all moments in time exist simultaneously, but they are ordered to create the illusion of an unfolding experience by some physical mechanism. This, in the language of relativity, may be connected to a hypersurface in a world that extends beyond spacetime. Death in such a scenario may be merely a phase change.

Keywords Time · Minkowski spacetime · Flow of time · Time as an illusion · Mathematics and reality

13.1 Introduction

A couple of years after Einstein formulated special relativity, Minkowski in a famous speech argued that time should be welded to space to form spacetime. The result is a hybrid measure of separation, or interval, commonly called the Minkowski metric. It is the basis of quantum mechanics. By extension to curved as opposed to flat spacetime, we obtain a more complicated expression for the interval, which is the basis of cosmology. However, between the small systems of quantum theory and the large ones of cosmology, there are numerous others which can be adequately described by Newtonian mechanics and also involve time. An ongoing debate, in both philosophy and physics, has to do with the nature of time in its various applications. Especially: are the various usages of time in physics and everyday life consistent with a unique definition for it? Alternatively: while time occurs in many guises, what is the most useful way to view it at a conceptual level? We hope in what follows to answer these and related questions by re-examining the argument – espoused by Einstein, Eddington, Hoyle, Ballard and others – that time is essentially a subjective *ordering* device.

P.S. Wesson (✉)
Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*, Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_13,

In doing this, it will be necessary to debunk certain myths about time, and to clarify statements that have been made about it. Certainly, time has been a puzzling concept throughout history. For example, Newton in his *Principia* (Scholium I), stated that "Absolute, true and mathematical time, of itself, and from its own nature, flows equably without relation to anything external, and by another name is called duration." This sentence is often quoted in the literature, and is widely regarded as being in opposition to the nature of time as embodied later in relativity. However, prior to that sentence, Newton also wrote about time and space that ". . . the common people conceive these quantities under no other notions but from the relation they bear to sensible objects." Thus Newton was aware that the "common" people in the 1700s held a view of time and other physical concepts which was essentially the same as the one used by Einstein, Minkowski, Poincaré and others in the 1900s as the basis for relativity.

As a property of relativity, it is unquestionably true that the time t can be considered as a physical dimension, on the same basis as our measures (x y z) of three-dimensional space. The result is spacetime. In this, the time part involves the product of t with the speed of light c, which essentially transforms the "distance" along the time axis to a length ct. Due to this, the interval is also a measure of which points are (or are not) in contact via the exchange of photons. Those particles with real interval can be in contact, while those with imaginary interval cannot be in contact.

This way of presenting Minkowski spacetime is conventional and familiar. However, it has a corollary which is not so familiar: particles with zero interval are *coincident* in 4D. Einstein realized this, and it is the basis of his definition of simultaneity. But it is not a situation which most people find easy to picture, so they decompose 4D spacetime into 3D space and 1D time, and visualize a photon propagating through x y z over time t. Eddington, the noted contemporary of Einstein, also appreciated the subjective nature of the situation just described, and went on to argue that much of what is called objective in physics is in fact subjective or invented. The speed of light was also commented on later by a few deep thinkers such as McCrea and Hoyle, who regarded it as a mere man-made constant. From the Eddington viewpoint, one can argue that the decomposition of 4D Minkowski spacetime into separate 3D and 1D parts is a subjective act, so that in effect the photon has been invented as a consequence of separating space and time.

Below, we will enlarge on the possibly subjective nature of physics, with an emphasis on the concept of time. We will in fact suggest that time is a subjective ordering device, used by humans to make sense of their world. Several workers have expressed this idea, including Einstein (1955 in [12]), Eddington [6, 7], Hoyle [13, 14], Ballard [1] and Wesson [24]. We hope to show that this approach makes scientific sense, and from a common-day perspective has certain comforts.

Such an approach is, however, somewhat radical. So to motivate it, we wish to give a critique of other, more mainstream views. This will be short, because good reviews of the nature of time are available by many workers including Gold [9], Davies [4], Whitrow [26], McCrea [18], Hawking [11], Landsberg [16], Zeh [28], Woodward [27] and Halpern and Wesson [10]. We will discuss contending views of the nature of time in Section 13.2, introduce what seems to be a better approach in Section 13.3, and expand on the implications of this in Section 13.4. Although it is not essential, it will become apparent that our new approach to time is most productive when the world is taken to have more dimensions than the four of spacetime, in accordance with modern physics.

13.2 Physics and the Flow of Time

The idea that time flows from the past to the future, and that the reason for this has something to do with the natural world, has become endemic to philosophy and physics. However, this idea is suspect. We will in this section examine briefly the three ways in which the direction of time's "arrow" is commonly connected with physical processes, and argue that they are all deficient. Quite apart from technical arguments, a little thought will show that a statement such as the "flow of time," despite being everyday usage, is close to nonsensical. For the phrase implies that time itself can be measured with respect to another quantity of the same kind. This might be given some rational basis in a multidimensional universe in which there is more than one time axis (see below); but the everyday usage implies measuring the change of a temporal quantity against itself, which is clearly a contradiction in terms. Such a sloppy use of words appears to be tolerated because there is a widespread belief that the subjective, unidirectional nature of time can be justified by more concrete, physical phenomena.

Entropy is a physical concept which figures in the laws of thermodynamics. Strictly speaking, it is a measure of the number of possible states of a physical system. But more specifically, it is a measure of the disorder in a system; and since disorder is observed to increase in most systems as they evolve, the growth of entropy is commonly taken as indicative of the passage of time. This connection was made by Eddington, who also commented on the inverse relationship between information and entropy [6, 7]. However, the connection has been carried to an unreasonable degree by some subsequent writers, who appear to believe that the passage of time is equivalent to the increase of entropy. That this is not so can be seen by a simple counter-argument: If it were true, each person could carry a badge that registered their entropy, and its measurement would correlate with the time on a local clock. This is clearly daft.

A more acceptable application of the notion of entropy might be found in the many-worlds interpretation of quantum mechanics. This was proposed by Everett [8], and supported as physically reasonable by De Witt [5]. In it, microscopic systems bifurcate, and so define the direction of the future. In principle, this approach is viable. However, the theory would be better couched in terms of a universe with more than the four dimensions of spacetime; and interest in the idea of many worlds appears to have lapsed, because there is no known way to validate or disprove their existence.

Another physical basis for the passage of time which has been much discussed concerns the use of so-called retarded potentials in electromagnetism. The connection is somewhat indirect, but can be illustrated by a simple case where light propagates from one point to another. (This is what happens when humans apprehend things by the sense of sight, and is also how most information is transmitted by modern technology.) Let the signal be emitted at point P and observed at point O, where the distance between them is d and is traversed at lightspeed c. Now Maxwell's equations, which govern the interaction, are symmetric in the time t. (We are assuming that the distance is small enough that ordinary three-dimensional space can be taken as Euclidean or flat.) However, in order to get the physics right, we have to use the electromagnetic potential not at time t but at the retarded time $(t - d/c)$. This is, of course, the time "corrected" for the travel lag from the point P of emission to the point O of observation. Such a procedure may appear logical; but it has been pointed out by many thinkers that it automatically introduces a time asymmetry into the problem (see [4] for an extensive review). The use of retarded potentials, while they agree with observations, is made even more puzzling by the fact that Maxwell's equations are equally valid if use is made instead of the "advanced" potentials defined at $(t + d/c)$. In short, the underlying theory treats negative and positive increments of time on the same footing, but the real world appears to prefer the solutions where the past evolves to the future. Studies have been made of the symmetric case, called Wheeler/Feynman electrodynamics, where both retarded and advanced potentials are allowed. One argument for why we do not experience the signals corresponding to the advanced potentials is that due to Hoyle and Narlikar [15]. They reasoned that the unwanted signals would be absorbed in certain types of cosmological model, leaving us with a universe which is apparently asymmetric between the past and future. This explanation is controversial, insofar as it appeals to unverified aspects of the large-scale cosmos. On the small scale, it appears that the need for retarded potentials in electrodynamics leads to a locally-defined arrow of time; though whether this is due to objective physical reasons, or to some subjective bias on our part, remains obscure.

The big bang offers yet another way of accounting for the arrow of time. According to Einstein's theory of general relativity, everything we observe came into existence in a singularity at a specific epoch, which supernova data fix at approximately 13×10^9 years before the present. This description is familiar to all, and carries with it the implication that the universe in a dynamical sense has a preferred direction of evolution. However, closer examination shows that it is really the recession of the galaxies from each other, rather than the big bang, which identifies the time-sense of the universe's evolution. This was understood by Bondi [3], who was one of the founders with Gold and Hoyle of the steady-state theory. In it, matter is continuously created and condenses to form new galaxies, whose average density is thereby maintained even though the whole system is expanding. While no longer regarded as a practical cosmology, the steady-state theory shows that it is the motions of galaxies which essentially defines a preferred direction for time, rather than the (still poorly understood) processes by which they may have formed after the big bang. Let us, in fact, temporarily forget about the latter event, and consider an ensemble of gravitating galaxies. Then there are in principle only three modes of evolution: expansion, contraction and being static. The last can be ruled out, because it is widely acknowledged that such a state, even if it existed, would be unstable and tip into one of the other two modes. We are thus lead to the realization that if the arrow of time is dictated by the dynamical evolution of the universe, its sense is given *a priori* by a 50/50 choice analogous to flipping a cosmic coin. That is, there is no dynamical reason for believing that events should go forward rather than backwards in time. In addition to this, there is also the problem that there is no known physical process which can transfer a cosmic effect on a lengthscale of 10^{18} cm down to a human one of order 10^2 cm. In order to circumvent this objection, it has been suggested that the humanly-perceived arrow of time is connected instead to smaller-scale astrophysics, such as the nucleosynthesis of elements that determines the evolution of the Sun. This process might, via the notion of entropy as discussed above, be connected to geophysical effects on the Earth, and so to the biology of its human inhabitants. But it is really obvious, when we pick apart the argument, that there is no discernable link between the mechanics of the evolving universe and the sense of the passage of time which is experienced by people.

The preceding issues, to do with entropy, electrodynamics and cosmology, have the unfortunate smell of speculation. Dispassionate thought reveals little convincing connection between the time coordinate used in physics and the concept of age as used in human biology. We can certainly *imagine* possible connections between physical and human time, as for example in *Einstein's Dreams* by Lightman [17]. There, the effects of relativity such as time dilation are described in sociological contexts. But, there is a large gap between the fluid manner in which time can be manipulated by the novelist and the rigid transformations of time permitted to the physicist. Indeed, while the physicist may be able to handle the "t" symbol in his equations with dexterity, he looks clumsy and strained when he attempts to extend his theories to the practicality of everyday existence. That is why the sayings about time by physicists mainly languish in obscurity, while those by philosophers have wider usage.

In the latter category, we can consider the statement of Marcel Proust: "The world was not created at the beginning of time. The world is created every day." This appears to dismiss the big bang, and by implication other parts of physics, as irrelevant to the human experience of time. However, it is more rewarding to consider statements like the foregoing as pointed challenges to the physicist. To be specific: Is there a view of "time" which is compatible with the rather narrow usage of the word in physics, and yet in agreement with the many ways in which the concept is experienced by people?

13.3 Time as a Subjective Ordering Device

The differing roles which time plays in physics and everyday life has led some workers to the conclusion that it is a subjective concept. Let us consider the following quotes:

Einstein (as reported by Hoffman): "For us believing physicists the distinction between past, present and future is only an illusion, even if a stubborn one."

Eddington: "General scientific considerations, favour the view that our feeling of the going on of time is a sensory impression; that is to say, it is as closely connected with stimuli from the physical world as the sensation of light is. Just as certain physical disturbances entering the brain cells via the optic nerves occasion the sensation of light, so a change of entropy ... occasions the sensation of time succession, the moment of greater entropy being *felt* to be the later."

Hoyle: "All moments of time exist together." "There is no such thing as 'waiting' for the future." "It could be that when we make subjective judgments we're using connections that are non-local ... there is a division, the world divides into two, into two completely disparate stacks of pigeon holes."

Ballard: "Think of the world as a simultaneous structure. Everything that's ever happened, all the events that *will* ever happen, are taking place together." "It's possible to imagine that everything is happening at once, all the events 'past' and 'future' which constitute the universe are taking place together. Perhaps our sense of time is a primitive mental structure that we inherited from our less intelligent forbears."

The preceding four opinions about time have an uncanny similarity, given that they apparently originate independently of each other. However, they are all compatible with Eddington's view of science, wherein certain concepts of physics are not so much discovered as invented (see [24] for a short review). The subjective nature of time is also compatible with current views of particle physics and cosmology, wherein several worlds exist alongside each other [5, 8, 19, 20, 25]. It is important to realize that there need not be anything mystical about this approach. For example, Hoyle considers a 4D world of the usual type with time and space coordinates t and $x\ y\ z$ which define a surface $\phi\,(t, xyz) = C$. Here C is a parameter which defines a subset of points in the world. Changing C changes the subset, and "We could be said to live our lives through changes of C." In other words, the life of a person can be regarded as the consequence of some mechanism which picks out sets of events for him to experience.

What such a mechanism might be is obscure. Hoyle speculated that the mechanism might involve known physical fields such as electromagnetism, which is the basis of human brain functions. It might plausibly involve quantum phenomena, amplified to macroscopic levels by the brain in the manner envisaged by Penrose [20]. However, while the precise mechanism is unknown, some progress can be made in a general way by noting that Hoyle's C-equation above is an example of what in relativity is known as a *hypersurface*. This is the relation one obtains when one cuts through a higher-dimensional manifold, defining thereby the usual 4D world we know as spacetime. It is in fact quite feasible that the Minkowski spacetime of our local experience is just a slice through a world of more than four dimensions.

In fact, higher dimensions are the currently popular way to unify gravity with the interactions of particle physics, and reviews of the subject are readily available (e.g., [25] from the physical side and [19] from the philosophical side). Since we are here mainly interested in the concept of time, let us concentrate on one exact solution of the theory for the simplest case when there is only a single extra dimension. (See [23] for a compendium of higher-dimensional solutions including the one examined here.) Let us augment the time (t) and the coordinates of Euclidean space

$(x\ y\ z)$ by an extra length (l). Then by solving the analog of Einstein's equations of general relativity in 5D, the interval between two nearby points can be written

$$dS^2 = l^2 dt^2 - l^2 \exp i\,(\omega t + k_x x)\,dx^2 - l^2 \exp i\,\left(\omega t + k_y y\right) dy^2 \\ -l^2 \exp i\,(\omega t + k_z z)\,dz^2 + L^2 dl^2. \tag{13.1}$$

Here ω is a frequency, k_x *etc*. are wave numbers and L measures the size of the extra dimension. This equation, while it may look complicated, has some very informative aspects: (a) it describes a wave, in which parts of what are commonly called space can come into and go out of existence; (b) it can be transformed by a change of coordinates to a flat manifold, so what looks like a space with structure is equivalent to one that is featureless; (c) the signature is $+---+$, so the extra coordinate acts like a second time. These properties allow of some inferences relevant to the present discussion: (a) even ordinary 3D space can be ephemeral; (b) a space may have structure which is not intrinsic but a result of how it is described; (c) there is no unique way to identify time.

This last property is striking. It means that in grand-unified theories for the forces of physics, the definition of time may be ambiguous. This classical result confirms the inference from quantum theory, where the statistical interaction of particles can lead to thermodynamic arrows of time for different parts of the universe which are different or even opposed [21,22]. It should be noted that the existence of more than one "time" is not confined to 5D relativity, but also occurs in other N-dimensional accounts such as string theory [2]. Indeed, there can in principle be many time-like coordinates in an N-dimensional metric.

In addition, the definition of time may be altered even in the standard 4D version of general relativity by a coordinate transformation. (This in quantum field theory is frequently called a gauge choice.) The reason is that Einstein's field equations are set up in terms of tensors, in order to ensure their applicability to any system of coordinates. This property, called covariance, is widely regarded as essential for any modern theory of physics. However, if we wish to have equations which are valid irrespective of how we choose the coordinates, then we perforce have to accept the fact that time and space are malleable. Indeed, covariance even allows us to *mix* the time and space labels. Given the principle of covariance, it is not hard to see why physicists have abandoned the unique time label of Newton, and replaced it by the ambiguous one of Einstein.

We are led to the realization that the concept of time is as much a puzzle to the physicist as it is to the philosopher. Paradoxically, the average person in the street probably feels more comfortable about the issue than those who attempt to analyse it. However, it is plausible that time in its different guises is a device used by people to organize their existence, and as such is at least partially subjective in character.

13.4 Mathematics and Reality

In the foregoing, we saw that several deep thinkers have arrived independently at a somewhat intriguing view of time. To paraphrase them: time is a stubborn illusion (Einstein), connected with human sensory impressions (Eddington), so that all moments of time exist together (Hoyle), with the division between past and future merely a holdover from our primitive ancestors (Ballard). Perhaps the most trenchant opinion is that of Hoyle [13], who summarizes the situation thus: "There's one thing quite certain in this business. The idea of time as a steady progression from past to future is wrong. I know very well we feel this way about it subjectively. But we're all victims of a confidence trick. If there's one thing we can be sure about in physics, it is that all times exist with equal reality"

This view of time can be put on a physical basis. We imagine that each person's experiences are a subset of points in spacetime, defined technically by a hypersurface in a higher-dimensional world, and that a person's life is represented by the evolution of this hypersurface. This is admittedly difficult to visualize. But we can think of existence as a vast ocean whose parts are all connected, but across which a wave runs, its breaking crest precipitating our experiences.

A mathematical model for a wave in five dimensions was actually considered in the preceding section as Eq. (13.1). It should be noted that there is nothing very special about the dimensionality, and that it is unclear how many dimensions are required to adequately explain all of known physics. The important thing is that if we set the interval to zero, to define a world whose parts are connected in higher dimensions, then we necessarily obtain the hypersurface which defines experience in the lower-dimensional world. It is interesting to note that the behaviour of that hypersurface depends critically on the number of plus and minus signs in the metric (i.e., on the signature). In the canonical extension of Einstein's theory of general relativity from four to five dimensions, the hypersurface has two possible behaviours. Let us express the hypersurface generally as a length, which depends on the interval of spacetime s, or equivalently on what physicists call the proper time (which is the time of everyday existence corrected to account for things like the motion). Then the two possible behaviours for the hypersurface may be written

$$l = l_o \exp(s/L) \text{ and } l = l_o \exp(is/L). \tag{13.2}$$

Here l_o is a fiducial value of the extra coordinate, L is the length which defines the size of the fifth dimension, and s is the aforementioned interval or proper time. The two noted behaviours describe, respectively, a growing mode and an oscillating mode. The difference between the two modes depends on the signature of the metric, and is indicated by the absence or presence of $i \equiv \sqrt{-1}$ in the usual manner. So far, the analysis follows the basic idea about experience due to Hoyle but expressed in the language of hypersurfaces as discussed by Wesson (see [14, 25]). However, it is possible to go further, and extend the analysis into the metaphysical domain for those so inclined. This by virtue of a change from the growing mode to the oscillatory mode, with the identification of the former with a person's material life and the latter with a person's spiritual life. That is, we obtain a simple model wherein existence is described by a hypersurface in a higher-dimensional world, with two modes of which one is growing and is identified with corporeal life, and one is wave-like and is identified with the soul, the two modes separated by an event which is commonly called death.

Whether one believes in a model like this which straddles physics and spirituality is up to the individual. (In this regard, the author is steadfastly neutral.) However, it is remarkable that such a model can even be formulated, bridging as it does realms of experience which have traditionally been viewed as immutably separate. Even if one stops part way through the above analysis, it is clear that the concept of time may well be an illusion. This in itself should be sufficient to comfort those who fear death, which should rather be viewed as a phase change than an endpoint.

13.5 Conclusion

Time is an exceptionally puzzling thing, because people experience it in different ways. It can be formalized, using the speed of light, as a coordinate on par with the coordinates of ordinary three-dimensional space. But while spacetime is an effective tool for the physicist, this treatment of time seems sterile to the average person, and does not explain the origin of time as a concept. There are shortcomings in purely physical explanations of time and its apparent flow, be they from entropy, many-worlds, electromagnetism or the big bang. Such things seem too abstract and remote to adequately explain the individual's everyday experience of time. Hence the suggestion that time is a subjective ordering device, invented by the human mind to make sense of its perceived world.

This idea, while not mainstream, has occurred to several thinkers. These include the philosopher Proust, the scientists Einstein, Eddington and Hoyle, and the novelist Ballard. It is noteworthy that the idea appears to have its genesis independently with these people. And while basically psychological in nature, it is compatible with certain approaches in physics, notably Penrose's suggestion that the human brain may be a kind of amplification organ for turning tiny, quantum-mechanical effects into measurable, macroscopic ones. The idea of time as an ordering device was given a basis in the physics of relativity by Hoyle, who however only sketched the issue, arguing that the movement of a hypersurface would effectively provide a model for the progress of a person's life. This approach can be considerably developed, as outlined above, if we assume that the experience-interface is related to a 4D hypersurface in a 5(or higher)D world. Then it is possible to write down an equation for the hypersurface, which can have an evolutionary and an oscillatory phase, which might (if a person is so inclined) be identified with the materialistic and spiritual modes of existence. Perhaps more importantly, in this 5D approach, the interval (or "separation") between points is zero, so all of the events in the world are in (5D) causal contact. In other words, everything is occurring simultaneously.

That this picture may be difficult to visualize just bolsters the need for something like the concept of time, which can organize the simultaneous sense data into a comprehensible order.

Time, viewed in this manner, is akin to the three measures of ordinary space, at least insofar as how the brain works. Humans have binocular vision, which enables them to judge distances. This is an evolutionary, biological trait. Certain other hunting animals, like wolves, share it. By comparison, a rabbit has eyes set into the sides of its head, so while it can react well to an image that might pose a threat, it cannot judge distance well. But even a human with good vision finds it increasingly difficult to judge the relative positions of objects at great distance: the world takes on a two-dimensional appearance, like a photograph, or a landscape painting. In the latter, a good artist will use differing degrees of shade and detail to give an impression of distance, as for example when depicting a series of hills and valleys which recede to the horizon. Likewise, the human brain uses subtle clues to do with illumination and resolution to form an opinion about the relative spacing of objects at a distance. This process is learned, and not perfectly understood by physiologists and psychologists; but is of course essential to the adequate functioning of an adult person in his or her environment. Astronomers have long been aware of the pitfalls of trying to assess the distances of remote objects. Traditionally, they measured offsets in longitude and latitude by means of two angles indicated by the telescope, called right ascension and declination. But they had no way of directly measuring the distances along the line of sight, and so referred to their essentially 2D maps as being drawn on the surface of an imaginary surface called the celestial sphere. Given such a limited way of mapping, it was very hard to decide if two galaxies seen close together on the sky were physically close or by chance juxtaposed along the line of sight. In lieu of a direct method of distance determination, astronomers fell back on probability arguments to decide (say) if two galaxies near to each other on a photographic plate were really tied together by gravity, or merely the result of a coincidental proximity in 2D while being widely separated in 3D. The situation changed drastically when technological advances made it easier to measure the redshifts of galaxies, since the redshift of a source could be connected via Hubble's law to the physical distance along the line of sight. Thus today, combining angular measurements for longitude and latitude with redshifts for outward distance, astronomers have fairly good 3D maps of the distribution of galaxies in deep space.

In effect, astronomers have managed to replace the photograph (which is essentially 2D) by the hologram (which provides information in 3D). However, whether this is done for a cluster of galaxies or a family portrait, the process of evaluating distance is a relatively complicated one. The human brain evaluates 3D separations routinely, and we are not usually aware of any conscious effort in doing so. But this apparently mundane process is also a complicated one. If we take it that the concept of time is similar to the concept of space, it is hardly surprising that the human brain has evolved its own subtle way of handling "separations" along the time axis of existence.

Thus the idea of time as a kind of subjective ordering device, by which we make sense of a simultaneous world, appears quite natural.

References

1. Ballard, J.G.: Myths of the Near Future. Triad-Panther, London (1984)
2. Bars, I., Deliduman, C., Minic, D.: Supersymmetric two-time Physics. Phys. Rev. D **59**, 125004 (1999)
3. Bondi, H.: Cosmology. Cambridge University Press, Cambridge (1952)
4. Davies, P.C.W.: The Physics of Time Asymmetry. University of California Press, Berkeley, CA (1974)
5. De Witt, B.S.: Quantum mechanics and reality. Phys. Today **23**(9), 30 (1970)
6. Eddington, A.S.: The Nature of the Physical World. Cambridge University Press, Cambridge (1928)
7. Eddington, A.S.: The Philosophy of Physical Science. Macmillan, New York (1939)
8. Everett, H.: 'Relative State' Formulation of quantum mechanics. Rev. Mod. Phys. **29**, 454 (1957)
9. Gold, T. (ed.): The Nature of Time. Cornell University Press, Ithaca, NY (1967)
10. Halpern, P., Wesson, P.S.: Brave New Universe. J. Henry Press, Washington, DC (2006)
11. Hawking, S.W.: A Brief History of Time. Bantam Press, New York (1988)
12. Hoffmann, B.: Albert Einstein, Creator and Rebel. New American Lib., New York (1972)
13. Hoyle, F.: October the First is Too Late. Fawcett-Crest, Greenwich, Conn (1966)
14. Hoyle, F., Hoyle, G.: Fifth Planet. Heinemann, London (1963)
15. Hoyle, F., Narlikar, J.V.: Action at a Distance in Physics and Cosmology. Freeman, San Francisco (1964)
16. Landsberg, P.T.: In: Sarlemijn, A., Sparnaay, M.J. (eds.): The Physical Concept of Time in the 20th Century. Physics in the Making, p. 131. Elsevier, Amsterdam (1989)
17. Lightman, A.: Einstein's Dreams. Random House, New York (1993)
18. McCrea, W.H.: Spontaneous Emission of Light in the Universe. Quart. J. Roy. Astron. Soc. **27**, 137 (1986)
19. Petkov, V. (ed.): Relativity and the Dimensionality of the World. Springer, Berlin (2007)
20. Penrose, R.: The Emperor's New Mind. Oxford University Press, Oxford (1989)
21. Schulman, L.S.: Time's Arrow and Quantum Measurement. Cambridge University Press, Cambridge (1997)
22. Schulman, L.S.: Time's arrows and quantum measurement. Phys. Rev. Lett. **85**, 897 (2000)
23. Wesson, P.S.: Space, Time, Matter. World Scientific, Singapore (1999)
24. Wesson, P.S.: Observatory 120 (1154), 59. *Ibid.*, 2001, 121 (1161), 82 (2000)
25. Wesson, P.S.: Five-Dimensional Physics. World Scientific, Singapore (2006)
26. Whitrow, G.J.: The Natural Philosophy of Time. Oxford University Press, Oxford (1980)
27. Woodward, J.F.: Making the Universe Safe for Historians: Time Travel and the Laws of Physics. Found. Phys. Lett. **8**, 1 (1995)
28. Zeh, H.-D.: The Physical Basis of Time. Springer, Berlin (1992)

Chapter 14
Consequences of Minkowski's Unification of Space and Time for a Philosophy of Nature

Herbert Pietschmann

Abstract Philosophy of nature is defined as opposed to the model of physics as well as to epistomology and transcendental philosophy. The question "what is space" is first dealt with in the old paradigm with space and time separated. The philosophical notion of "aporon" is introduced as the self-contradictory unity of object and communication. It is argued, that gravitation is the necessary communication providing the "existence" of space. The concept is extended to Minkowski's space-time. In order to achieve this, the philosophical notion of "now" is introduced and the problem of "duration" or "substance" is dealt with.

Keywords Philosophy of nature · Space · Time · Aporon · Existence

14.1 Philosophy of Nature

For the following considerations it is imperative to carefully distinguish "philosophy of nature" from the scientific model as well as from epistemological or transcendental questions. Philosophy of nature is based on the fundamental realization which Carl Friedrich von Weizsäcker has phrased clearly[1]:

> The relation of philosophy and so-called positiv science can be formulated as follows: The condition for success of the scientific method was to renounce those questions which are asked by philosophy. With that it is claimed that the success of science rests on the renouncement of certain questions. Among those are in particular the basic questions of the considered field.

H. Pietschmann (✉)
Emeritus at Faculty of Physics, University of Vienna, Austria

[1] C.F. von Weizsäcker: Deutlichkeit. Hanser Verlag, München (1978), p. 167. (Original: Das Verhältnis der Philosophie zur so genannten positiven Wissenschaft lässt sich auf die Formel bringen: Philosophie stellt diejenigen Fragen, die nicht gestellt zu haben die Erfolgsbedingung des wissenschaftlichen Verfahrens war. Damit ist also behauptet, dass die Wissenschaft ihren Erfolg unter anderem dem Verzicht auf das Stellen gewisser Fragen verdankt. Diese sind insbesondere die eigenen Grundfragen des jeweiligen Faches.)

V. Petkov (ed.), *Minkowski Spacetime: A Hundred Years Later*,
Fundamental Theories of Physics 165, DOI 10.1007/978-90-481-3475-5_14,

To illustrate this statement let us recall, that Newton did *not define* "time, space, place, and motion, for they are known to everybody".[2] Rather, he *distinguished* between, on the one hand, the relative, apparent, common conception of them, and, on the other hand, the absolute, true, mathematical quantities themselves. Thus, he distinguished "absolute, true, and mathematical space" (which remains similar and immovable without relation to anything external) from "relative spaces" (which are measures of absolute space defined with reference to some system of bodies or another, and thus a relative space may, and likely will, be in motion). And he distinguished "absolute, true, and mathematical time" from "relative, seeming and ordinary time." He did not ask questions as to what space and time *is in itself*!

Likewise, Richard Feynman in his famous lectures on physics writes[3]:

> What *is* time? It would be nice if we could find a good definition of time. ... Maybe it is just as well if we face the fact that time is one of the things we probably cannot define (in the dictionary sense). ... What really matters anyway is not how we *define* time, but how we measure it.

Thus the particular properties of the mathematical model constructed by physics shall not enter our considerations as long as we are interested in philosophy of nature. However, some basic insights which go beyond pure quantitative description can arouse considerations also valuable for philosophy of science. (An example important for this work is the unification of space and time in the mathematical model, which cannot be ignored by philosophy of space and time!)

In a sense, this worked both ways when Einstein found out that absolute time cannot be measured in principle, from which he concluded that it should not even enter the mathematical model.

I shall follow the differentiation of Andrew Ward who writes[4]:

> Three views of space and time are on offer here; According to the first, the Newtonian (or absolute) view, space and time exist not only independently of being perceived, but independently of any objects (understood as *things in themselves*) in space and time. This is the view being referred to when it is asked if space and time are real existences. According to the second, Leibnizian (or relational), view, space and time do exist independently of being perceived, but do not exist independently of things in themselves. Space and time are merely the relations holding between things in themselves, which we confusedly perceive by means of sensations in our minds. This is the view that is being referred to when it is asked if space and time are only relations or determinations of (things in themselves), yet such as would belong to things (in themselves) even if they were not intuited.
>
> Kant rejects *both* these views in favour of a third: viz. That space and time are to be equated with our outer and inner a priori intuitions, respectively.

Here, philosophy of nature (space and time), shall be based on the relational (or Leibnizian) view. In contrast, Kant is interested in transcendental idealism and defines:

---

[2] Heuser, H.: Der Physiker Gottes – Isaac Newton oder die Revolution des Denkens, p. 118. Herder Verlag, Freiburg (2005).

[3] Feynman, R.P., Leighton, R.B., Sands, M.: The Feynman Lectures on Physics, vol. I, p. 5–10. Addison-Wesley, Reading MA (1963).

[4] Ward, A.: Kant – The Three Critiques, p. 33. Polity Press, Cambridge (2006).

> I entitle *transcendental* all knowledge which is occupied not so much with objects as with the mode of our knowledge of objects, insofar as this mode of knowledge is meant to be possible a priory.[5]

After the change of paradigm by Einsteins Relativity, philosophy of space and time can no longer restrict itself to Kant's extraordinary analysis which is based on the idea of synthetic judgements a priory. Einstein has pointed this out very clearly[6]:

> As far as the laws of mathematics refer to reality, they are not certain; and so far as they are certain, they do not refer to reality.

14.2 What Is Space?

The question "what is space" is one of the questions to which Carl Friedrich von Weizsäcker refers in his statement above. The notion of space is intrinsically contradictory (in the sense of antinomy or απωρια), for space is simultaneously neither something nor nothing. If it were something, we would call it object and not space. If it were nothing, where would objects be (other than in space)?

Within the mathematical model, this contradiction can neither be eliminated, nor brought to a dialectic synthesis. It suffices to find what I call "operational mastering".[7] This means, that we find an algorithm that allows to handle the situation in an intersubjective way without falling back to the elimination of the contradiction by either-or. In my view, the situation has been operationally mastered by means of the coordinate system (and later mathematical constructions superseding the original notion of coordinate system).[8]

Let us consider this in more detail. In order to secure the place of an object (represented by a point, e.g., its center of gravity), we need a coordinate system and three numbers (the coordinates, x_1, x_2, x_3, say). But these three numbers are irrelevant, because the coordinate system can be rotated and/or displaced at our arbitrary convenience without changing the physics we want to describe. Relevance appears only when we have two distinct points, i.e., six coordinates, x_1, x_2, x_3, and y_1, y_2, y_3, say. Only one number is of physical relevance: the distance l between the two points, with $l^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2$. (Obviously, we can transform the coordinate system so that we get $y_1 = l,\ y_2 = y_3 = x_1 = x_2 = x_3 = 0$).

The operational mastering of the intrinsic contradiction of space succeeds by recognizing that we need a coordinate system, but only those numbers are of physical relevance which remain *invariant* under an arbitrary transformation of the

---

[5] Kant, I.: Critique of Pure Reason, 2nd edn, p. 25 (1787).

[6] A. Ward: loc. cit. p. 25.

[7] Pietschmann, H.: Phänomenologie der Naturwissenschaft, 2nd edn, p. 56f. Ibera/European University Press, Wien (2007).

[8] Pietschmann, H.: Abhandlungen der Humboldt-Ges. f. Wissenschaft, Kunst, Bildung, p. 3–6. Fischer, H., Haberland, D. (eds.) vol. **18** (2003).

coordinate system. The intrinsic contradiction is reflected when we observe that *the coordinate system is always simultaneously necessary and irrelevant!*

Can philosophy of nature deal with this contradiction in a deeper way?

At the University of Vienna, since more than 40 years exists a "Philosophical-Scientific Working Group" interested in philosophy of science. It has taken up this question and came to the following answer:[9] "Space is the condition for the possibility that an object can be external to another object."

As long as we are thinking in terms of appearances, this is still a transcendental observation, i.e., it is a statement about our mind rather than about space in itself. In our relational view, "being external" has to be communicated between the objects in themselves! It is my ansatz, that this communication is the gravitational interaction between the objects!

14.3 Atom, Monad and Aporon

The notion of ατωμως (indivisible and indestructible atom) introduced by Democrit has been shifted from the physical atom to the fundamental particles (leptons, quarks, gauge-bosons and possibly Higgs). In spite of the fact that they are neither indivisible nor indestructible, they are considered to be the basic constituents of matter in the physical model for they do not have sub-constituents into which they can be divided. (They are divisible into other constituents of the same level, not into sub-constituents!)

In quantum field theory we have learned, that constituents ("bare particles") cannot be separated from their interactions. As soon as we add interactions, the constituents change from "bare particles" to "dresses or physical particles" (renormalization procedure). Only the latter can be observed in experiments. (In a sense, this may be an *analogy* – not a correspondence! – to Kant's thing in itself and its appearance.)

From quantum mechanics we have learned, that constituents may only be considered individuals as long as they are isolated. Whenever they get in contact with similar constituents, the typical quantum mechanical phenomenon of entanglement may occur (Erwin Schrödinger spoke of the "antinomies of entanglement"). Say two identical particles lose their individuality and become an entangled pair (which I prefer to call "double-particle"[10]) Schrödinger said[11]: "*When two systems interact, it is not so that their ψ-functions interact, rather they cease to exist and a unique one for the whole system replaces them.*"

We shall ask the question what that means for a philosophy of nature.

[9] Schwarz, G.: Raum und Zeit als naturphilosophisches Problem. WUV Universitätsverlag Wien (1992). (first edition: Herder Verlag, Wien, 1972).

[10] Pietschmann, H.: Quantenmechanik verstehen, p. 105. Springer-Verlag, Berlin (2001).

[11] Schrödinger, E.: Die Naturwissenschaften **23** (1935) 807, 823, 844. § 15.

Let me start by mentioning a philosophical approach to constituents: The monad-doctrine of Leibniz. The monad of Leibniz is indivisible in the sense that any division destrois the original monad. Thus the notion of monad is capable of embracing the (psychological) "I" and even large objects like stars or planets.[12] However, there is a notorious problem: "Monads do not have windows through which something may enter or leave"[13] and thus any communication between them seems impossible. In order to explain it, Leibniz had to invoke the principle of "preestablished harmony"[14] which, to me, seems a bit unsatisfactory. (It is interesting to remember that Albert Einstein has referred to Leibnizens concept of "preestablished harmony" when he tried to explain the practical uniqueness of theoretical descriptions in view of the fact that there is no logical path from perceptions to theoretical principles.[15])

In order to reconcile the philosophical concept of monad with their non-deniable communications, I am compelled to take the following antinomy into my definition:

Thesis: Monads are windowless.

Antithesis: Monads communicate with each other (as well as with themselves).

Thus I start from the very beginning with a different notion, which comprises both sides of the antinomy, the idea of basic unity as well as the idea of compulsory communication. I call these elements of being "aporon" after the greek απωρια. Needless to say that I was guided by the achievements of quantum field theory and quantum mechanics (see above) without sticking to the necessarily reductionist view of the physical model. I was also encouraged to find out that the Japanese philosopher Nishida starts from similar thoughts; his basic concept is the "absolut contradictory self-identity".[16] (Let me mention in passing that Nishida also commented Minkowski by writing:[17] "*Some people condemn Minkowski's Theory of space-time as abstract and difficult to imagine. But I am not of this opinion...*").

In our relational view we can now avoid a backlash into transcendental idealism by amending the above statement about space in the following way: "Space is the condition for the possibility that an aporon can be external to another aporon where 'being external' is communicated between the aporons by gravitation."

14.4 The Velocity of Communication

We know from Einsteins special theory of relativity that the vacuum velocity of light (c) is the limiting speed for any physical object and the maximum velocity for transmission of information (communication). It is therefore often called fundamen-

[12] See e.g.. Klein, H.D.: Vernunft und Wirklichkeit, Band 2. Oldenburg Verlag Wien (1975) p. 85f.

[13] Leibniz: Monad-doctrine, § 7.

[14] Leibniz: Monad-doctrine, § 78.

[15] Einstein, A.: Mein Weltbild, Querido Publishers Amsterdam (1934) p. 109.

[16] I am indebted to Hashi Hisaki for pointing this out to me.

[17] Quoted from Hashi, H.: The Significance of Einstein's Theory of Relativity in Nishida's "Logic of Field". Philosophy East & West, vol. 57(4), Oct. 2007, p. 469.

tal velocity. Indeed, the fundamental velocity is the bridge between space and time ($x_0 = c.t$ in the model) and can therefore NOT be measured! Instead, its numerical value *defines* (arbitrarily) the relation of units of time and of length. Particle physicists simply set $c = 1$, for practical purposes the International Union of Pure and Applied Physics, after defining the second as the time unit, states:

> The metre is the length of the path travelled by light in vacuum during a time interval of 1/299792458 of a second.

The number is purposefully chosen such that for ordinary life we do not have to change our meter rods. Whoever thinks he or she can measure the speed of light simply gauges his or her measuring rods instead.

The fundamental velocity is the velocity of communication between aporons. (Just as the velocity of light may be reduced in a medium, the velocity of communication may be reduced by the masses of intermediary bosons; these are secondary effects of no importance here.)

Since the communication velocity is finite, each aporon communicates with any other aporon in the latter's past. (If an omnipotent being could remove the sun at an instant, the earth would stay in its original path for about 8 more minutes before it disappears on a straight line in the depth of the universe.)

14.5 Space-Time Instead of Space and Time

Hermann Minkowski opened his famous talk at the 80th assembly of the "Gesellschaft Deutscher Naturforscher und Ärzte" on September 21, 1908, in Cöln as follows[18]:

> Gentlemen! The conceptions on space and time which I am going to develop for you have grown on experimental physical ground. That is their strength. Their tendency is radical. From now on, space per se and time per se will completely turn into shadow, and only a kind of union of both shall live on.

It is imperative to accept that space and time have been *unified* but not *identified!* Space and time are no longer *separate*, but they remain to be *different*!

Since space and time do not have any separate meaning any more, we have to adapt the above statement about space. By communication, aporons are no longer external to each other in space, but in space-time. Consequently, one and the same aporon is external to itself at different moments of time. This opens the age-old question of "substance", i.e., the question how an object can stay identical to itself in spite of the unavoidable change in time.

[18] Published in Fortschritte der mathematischen Wissenschaften, Heft 2. Leipzig 1915, pp. 56–68. (Original: Meine Herren! Die Anschauungen über Raum und Zeit, die ich Ihnen entwickeln möchte, sind auf experimentell physikalischem Boden erwachsen. Darin liegt ihre Stärke. Ihre Tendenz ist eine radikale. Von Stund' an sollen Raum für sich und Zeit für sich völlig zu Schatten herabsinken, und nur noch eine Art Union der beiden soll Selbstständigkeit bewahren.)

Any philosophy of nature has to give an answer to this fundamental problem. For Plato it was the "idea," for Aristotle the "usia" and Kant relies on the continuous nature of time[19]:

> If we assume that something absolutely begins to be, we must have a point of time in which it was not. But to what are we to attach this point, if not to that which already exists? For a preceeding empty time is not an object of perception.

In the physical model, that is in quantum field theory, the problem sheds its shadow also. Any particle can always annihilate with a virtual antiparticle which has been pair-created before within the time interval allowed by the uncertainty relation. Thus it travels on as a seemingly different particle if we do not refer to its "aporon nature" as described above (two entangled particles are not to be considered different!). This situation led R.P. Feynman to the illustrative description in one of his Nobel prize papers[20]:

> It is as though a bombardier flying low over a road suddenly sees three roads and it is only when two of them come together and disappear again that he realizes that he has simply passed over a long switchback in a single road.

In order to approach this problem in philosophy of nature, we first have to elaborate on the unity and difference of space and time.

At any point in space-time (in the physical model called "event"), the separation between space and time is no problem. However, it is not invariant because for a relatively moving event, the separation is different. Thus, space and time cannot *invariantly* be distinguished. But we learn from the model of physics (i.e., special relativity theory), that there is a new, invariant distinction: Spacelike and timelike pairs of events! This is yet another corroboration from the model for my notion of aporon. Event-pairs are necessary for the new, invariant distinction between spacelike and timelike!

For spacelike separated events the order of time is not fixed, i.e., A can be either before or after B, depending on the state of motion of an observer! It is interesting to observe the analogy (not correspondence!) to Kant's distinction between space and time[21]:

> Kant's reply is that it is only in so far as I recognize that... I could have apprehended the successive representations in the *reverse* order that the experience of their coexistence can arise. ... In other words, I need to recognize that it *must* have been possible for me... to have apprehended the representations in the reverse order. The experience of an objective coexistence is made possible through the recognition of the *reversibility* of apprehension.

In the physical model, this important distinction is of course quantitative, i.e., it depends on the sign of an invariant.[22] In a philosophy of nature, the distinction

---

[19] Kant, I.: Critique of Pure Reason, 2nd edition, p. 231 (1787).

[20] Feynman, R.P.: The Theory of Positrons. Phys. Rew. **76**, p. 749 (1949).

[21] A. Ward, loc.cit., p. 75. (Emphasis original.)

[22] For two events (ct, x_1, x_2, x_3) and (ct', y_1, y_2, y_3) the invariant is $s^2 = [c(t-t')]^2 - [(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2]$. $s^2 > 0$ for time-like pairs, $s^2 < 0$ for space-like pairs.

must be qualitative rather than quantitative. Aporon is the unity of object and communication, but not any pair of aporons can communicate. Time is generated by the communication between aporons. (Communication between two aporons is only possible when they are in timelike separation.) Space cannot be thought of as "present" at any given time, since different spacelike separated aporons are not in communication. (When we look at a mirror, we do not observe ourselves in our present but rather a tiny fraction of time earlier, since the light takes this fraction of time to travel to the mirror and back into our eyes.)

Since each aporon communicates with any other aporon in the latter's past, this defines a "NOW" for an aporon ("now" – German "Jetzt" – is for an aporon what "event" is for a particle). It is the manifold of all aporons with whom the aporon in question communicates. Let me give an anthropomorphic example again: For a human being, "now" is everything in its immediate neighbourhood with which it communicates, but also sun, moon and stars. Our human being receives communications (light, warmth or any other impressions) from the moon as it was just over a second before, from the sun as it was about 8 min before and from stars as they were may be even millions of years before. All that is "now". We can therefore extend the philosophical description of space to space-time: "Space-time is the condition for the possibility that a now can be external to another now."

As stated above, this requires an answer to the problem of continuation or identity ("substance"). In this picture, self-identity is not guaranteed by an Idea, Usia or continuation of time. Two rows in time are as separate as in space. So how can an aporon be identified with itself in the past? It is not substance, it is self-communication which re-creates the aporon from itself at any moment. (In the model of physics, i.e., quantum field theory, the corresponding technical term is "self-energy" created by self-interaction!)

In the case of an aporon which is a celestial body, it is quite obvious that self-gravitation is at the root of its unity. A deeper analysis of this perennial problem requires the inclusion of the other interactions as well, but that goes beyond the scope of this article.