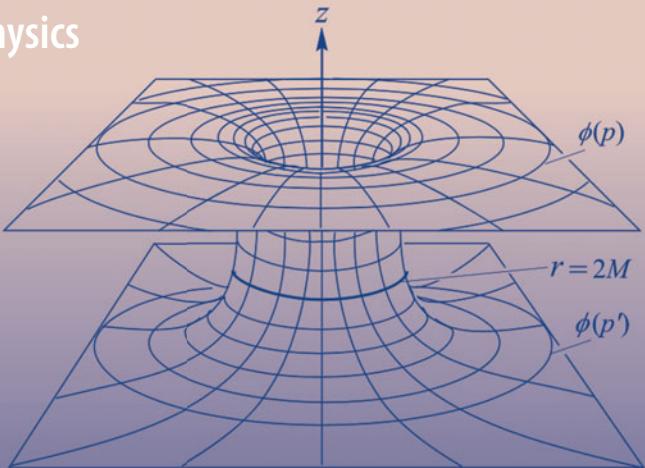


Graduate Texts in Physics

Canbin Liang
Bin Zhou



Differential Geometry and General Relativity

Volume 1

Translated and Revised by
Weizhen Jia and Bin Zhou



Science Press
Beijing



Springer

Graduate Texts in Physics

Series Editors

Kurt H. Becker, NYU Polytechnic School of Engineering, Brooklyn, NY, USA

Jean-Marc Di Meglio, Matière et Systèmes Complexes, Bâtiment Condorcet,
Université Paris Diderot, Paris, France

Sadri Hassani, Department of Physics, Illinois State University, Normal, IL,
USA

Morten Hjorth-Jensen, Department of Physics, Blindern, University of Oslo, Oslo,
Norway

Bill Munro, NTT Basic Research Laboratories, Atsugi, Japan

Richard Needs, Cavendish Laboratory, University of Cambridge, Cambridge, UK

William T. Rhodes, Department of Computer and Electrical Engineering and
Computer Science, Florida Atlantic University, Boca Raton, FL, USA

Susan Scott, Australian National University, Acton, Australia

H. Eugene Stanley, Center for Polymer Studies, Physics Department, Boston
University, Boston, MA, USA

Martin Stutzmann, Walter Schottky Institute, Technical University of Munich,
Garching, Germany

Andreas Wipf, Institute of Theoretical Physics, Friedrich-Schiller-University Jena,
Jena, Germany

Graduate Texts in Physics publishes core learning/teaching material for graduate- and advanced-level undergraduate courses on topics of current and emerging fields within physics, both pure and applied. These textbooks serve students at the MS- or PhD-level and their instructors as comprehensive sources of principles, definitions, derivations, experiments and applications (as relevant) for their mastery and teaching, respectively. International in scope and relevance, the textbooks correspond to course syllabi sufficiently to serve as required reading. Their didactic style, comprehensiveness and coverage of fundamental material also make them suitable as introductions or references for scientists entering, or requiring timely knowledge of, a research field.

Canbin Liang · Bin Zhou

Differential Geometry and General Relativity

Volume 1

Canbin Liang
Department of Physics
Beijing Normal University
Beijing, China

Bin Zhou
Department of Physics
Beijing Normal University
Beijing, China

Translated by

Weizhen Jia 
Department of Physics
University of Illinois Urbana-Champaign
Urbana, IL, USA

Bin Zhou
Department of Physics
Beijing Normal University
Beijing, China

ISSN 1868-4513

ISSN 1868-4521 (electronic)

Graduate Texts in Physics

ISBN 978-981-99-0021-3

ISBN 978-981-99-0022-0 (eBook)

<https://doi.org/10.1007/978-981-99-0022-0>

Jointly published with Science Press

The print edition is not for sale in China (Mainland). Customers from China (Mainland) please order the print book from: Science Press.

Translation from the Chinese Simplified language edition: “微分几何与广义相对论/Wei fen ji he yu guang yi xiang dui lun” by Canbin Liang and Bin Zhou, © Science Press 2006. Published by Science Press. All Rights Reserved.

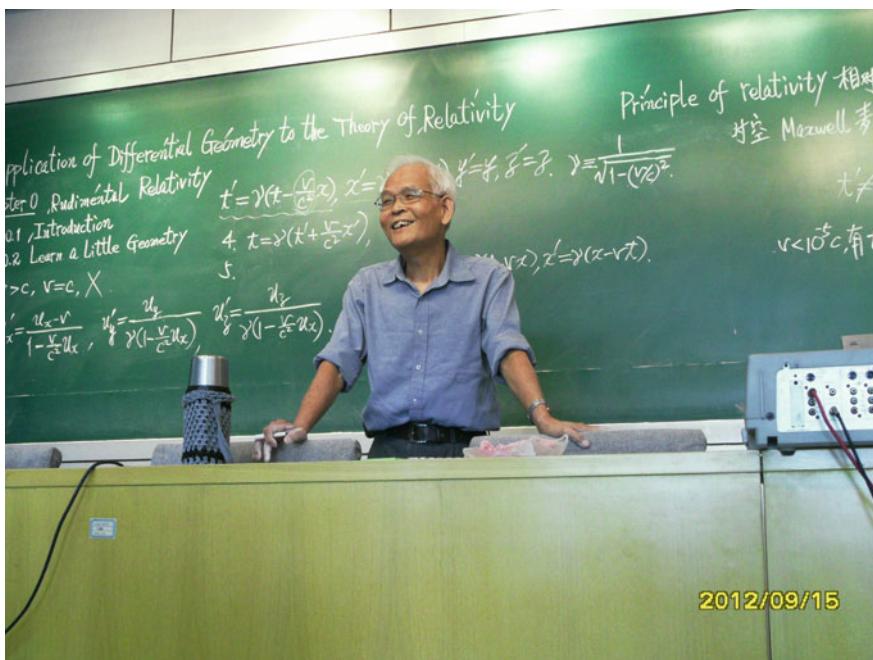
© Science Press 2023

This work is subject to copyright. All rights are reserved by the Publishers, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publishers, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publishers nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publishers remain neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721,
Singapore



Professor Canbin Liang (1938–2022) giving a lecture in 2012, photograph by Gui-Rong Liang

Foreword

I first met Canbin Liang during a visit I made to China in 1980. Liang had just received authorization from the Chinese government to come to the United States for two years as a visiting scholar, and he asked if I could serve as his advisor. I agreed, and Liang spent 1981–83 in my group at the University of Chicago.

At that time, China had just emerged from the Cultural Revolution, so, as might be expected, Liang had very little knowledge of modern developments in general relativity when he arrived. But he had an enormous interest in acquiring a deep understanding of the foundations of general relativity and a tremendous dedication to doing so. At the time, I was in the midst of writing my *General Relativity* book and was giving considerable thought as to how the subject should be presented. Liang carefully read drafts of my book and we had many discussions of issues in general relativity. Liang also had many interactions with Bob Geroch during his years in Chicago.

After he returned to China, Liang and I maintained our friendship and we continued to correspond on many issues in general relativity. Liang devoted himself to teaching the modern approach to general relativity to students in China and he published his book *Differential Geometry and General Relativity* in Chinese. I learned from a number of students who came from China to our Ph.D. program at the University of Chicago that this book—as well as Liang’s renown lecture course in general relativity—was instrumental in introducing generations of Chinese students to the modern approach to general relativity.

Liang’s book is quite similar to my *General Relativity* book in its coverage and presentation of the core material. However, while my treatment of the mathematical material focuses on presenting the general, abstract results in as concise a way as possible, Liang’s book provides many simple examples that illustrate these results. Liang also makes considerable effort to warn readers of possible pitfalls that can arise in their understanding of the material. I believe that many readers will find these examples and additional discussion quite helpful for working through the meaning of the general concepts.

I am very pleased that *Differential Geometry and General Relativity* has been translated into English, so that students outside of China now also can benefit from Liang's insights and pedagogy.

January 2023

Robert Wald
Charles H. Swift Distinguished Service
Professor of Physics
University of Chicago
Chicago, USA

Preface to the English Edition

The present book is translated and revised based on the second edition of the original Chinese text under the guidance of the first author, the late Prof. Canbin Liang. As one of the most popular Chinese books on theoretical physics, this work has already influenced generations of Chinese physicists since it first came out in 2000. Now we are so pleased that this work is translated into English and can be accessed by a broader group of readers. As can be told by the Chinese readers, the writing style of Prof. Liang is quite distinctive, as he sometimes uses idioms, or even self-created expressions, in order to make the text more vivid and intuitive to readers. Although this brings certain difficulties to the translation, we have made our best attempt to keep the style of the original text, so that this translation can be not only faithful to but also as expressive as the original work.

Apart from translation, we also implemented a sizable amount of revision to this work. From minor typo fixes to major content updates, every single chapter has been improved to some extent. Some revision was based on the personal notes of Prof. Liang himself, as he has been organizing possible improvements to this work based on his teaching experience over the years. Besides this, the parts that have been heavily modified are primarily those on gravitational radiation (Sect. 7.9) and cosmology (Chap. 10), due to the rapid developments in these areas over the past two decades. For Sect. 7.9, we expanded the discussions on the gauge conditions and gravitational plane waves substantially. We also replaced the out-of-date introduction on the detection of gravitational waves with a new one, which includes the discussions on the interferometric detection and the recent progress since the first direct observation by LIGO. For Chap. 10, we decided to revise and upgrade the content of the original chapter comprehensively, and it is now divided into two parts. The first half is the new Chap. 10, which sets up the geometric foundation for cosmology, and focuses on the standard cosmological model. Particularly, we enhanced our mathematical descriptions for the spatial geometries of the universe and updated the observational data to the latest ones. The second half will introduce the (currently in development) “new standard cosmological model”, which includes inflation, dark matter and dark energy. This part will be presented in Volume II along with other advanced topics.

As has been mentioned in the prefaces of the Chinese editions, this work was influenced by many other textbooks, especially the classic text *General relativity* by Robert Wald. Many discussions in this work can be viewed as extensions of those concise and incisive lines in Wald's book, making them more accessible to beginners. However, the aim of this work is in no way to be a replacement for Wald's book. In fact, we encourage readers to refer to his book after reading this text. As Prof. Liang said, just like Wald's book does a great job of paving the way for readers to understand *The Large Scale Structure of Space-Time*, the masterpiece by Stephen Hawking and George Ellis, one of the goals of this work is to pave the way for reading Wald's book (or one at that level). In addition, the material in these books is also complementary in many ways. We hope this work, especially Volume I, can be an initiation for the beginners to differential geometry and general relativity, and can open up the world for them to absorb further knowledge from other great works.

We would like to thank those who helped us and contributed to this work. First, we express special thanks to Prof. James Nester, who has supported the preparation of this English edition since day one. He read the translated manuscript of this entire book carefully, and provided not just English refinements, but also plenty of professional comments, which notably improved the quality of this book. We also benefited a lot from discussions with him. The translator Weizhen Jia would like to thank his friend Brandon Buncher for his constant support. He read a large portion of the translated manuscript and always offered nice suggestions when Jia consulted him. In particular, he helped a lot with those expressions used by Prof. Liang that are hard to translate, which makes this work more accessible to readers from Western (and other non-Chinese) backgrounds. We would also like to sincerely thank Prof. Zhoujian Cao, who provided valuable suggestions on the manuscript of Sect. 7.9, and has always been very supportive to us during the preparation of this work. We also appreciate Prof. Bin Hu for his comments and suggestions on the manuscript of Chap. 10.

We would like to thank Nick Abboud and Marcus Rosales for their helpful comments, and Jinhuan He for providing a translation draft for part of Chap. 9. In addition, we thank Prof. Robert Wald for his support and the lovely foreword to the English edition. We would also like to express thanks to Prof. Jerzy Lewandowski, Prof. Zheng Zhao, Prof. Sijie Gao, Prof. Younge Ma and Prof. Tong-Jie Zhang for their support of the publication of this work, and to Dr. Mengchu Huang from Springer Nature for his assistance. We are also grateful to Ms. Jinxing Zhang for her generous help.

Prof. Liang was an extraordinary educator who dedicated his entire life to passing on the knowledge of physics to the next generation. At age 84, he was still giving lectures even just a few days before he became critically ill. He left us forever soon after that, just when this volume was about to be finalized; it is unfortunate that he could not be here in person to see it being published. We hope that the outcome of

this effort can benefit more readers around the world, fulfilling a wish of our beloved and highly esteemed professor.

January 2023

Weizhen Jia
University of Illinois at Urbana-Champaign
Urbana, USA

Bin Zhou
Beijing Normal University
Beijing, China

Preface to the Second Edition

Since its publication in 2000, the first edition of this book (the first volume) has attracted attention and received praise in the field of theoretical physics, especially in general relativity, in China. Due to a small print run, it was sold out in only 2 years. After its publication, I have been using this book as a textbook; thus far, I have used it five times to teach graduates and undergraduates. In addition to the teaching at Beijing Normal University, I was also invited to give lectures to both undergraduates in the Fundamental Science Class of Tsinghua University and graduate students from the Academy of Mathematics and Systems Science at the Chinese Academy of Sciences (CAS). The classes at the CAS have also attracted dozens of graduates and undergraduates from other departments of the CAS and from another 11 institutions of higher learning, including Peking University and Tsinghua University. I realized the importance of this work in promoting the subject, and at the same time, I also found that there are some mistakes and deficiencies in this work. Needing to improve the content and to supplement the work with more details, I set out to write a draft of the second edition. While creating the new draft, I consulted with many of my colleagues and students, including (in alphabetical order by their last names) Zhoujian Cao, Muxin Han, Zhiqian Kuang, Yongge Ma, Zhi Wang, Xiaoning Wu, Xuejun Yang, Hao Zhang, Hongbao Zhang, Bin Zhou and Meike Zhou. Among them, Zhoujian Cao, Zhiqian Kuang, Hongbao Zhang and Bin Zhou have made outstanding contributions. Through many discussions with Dr. Bin Zhou, I found that he not only has a vast amount of knowledge of mathematics and physics but also a clear and logical way of thinking. He always has a relatively clear, deep and accurate understanding of the mathematical and physical problems within and beyond this work. In this way, he is truly a rare and outstanding physicist. In order to further improve the writing quality, I decided to invite him to revise this work as the second author, and he agreed to my request. The close collaboration in the past 5 months has proved that this was the correct decision. I think Dr. Bin Zhou has indeed made remarkable contributions to the revision.

I would like to express special thanks to two other friends who helped me with and contributed to the writing of this work. The first one is Dr. Robert Wald, a professor at the University of Chicago and a member of the National Academy of Sciences

(NAS). He is an excellent advisor who enlightened me and offered generous help to my teaching and writing after I returned to China. The other is Researcher Zhiqian Kuang from the Institute of Mathematics, CAS. He reviewed many chapters of this work, putting forward many valuable suggestions. In addition, his profound thinking and deep understanding always benefited me when I discussed with him.

This work is divided into two volumes 1,¹ and their contents have been fully introduced in the preface of the first edition. The revision is not only a comprehensive rewrite of the first edition but also supplements the first edition with new content. The main supplements to the first volume include the Vaidya metric and the Kinnersley metric, conjugate points, embedding diagrams and dark energy. The main additions to the second volume include fiber bundles and their applications in physics, spaces of constant curvature, and the de Sitter and anti-de Sitter space times. Although the second edition contains much more difficult material, it still maintains the writing style of the first edition, that is, it is made to be as understandable as possible. In particular, the introductory parts of this work are designed to be accessible to beginners. The whole work can be used as a text for a graduate course and a reference for relativists as well. The first volume can also be used as a reference book for undergraduates who are in the second year or above. Physicists who are not in the field of relativity can also take the first five chapters of the first volume and some other chapters in the whole work, such as those on Lie groups and Lie algebras and fiber bundle theory, as an introduction to differential geometry.

As to the writing style, another feature of this work is that it contains two parts—compulsory reading and optional reading—to meet the needs of readers at different levels. There are a large number of exercises in each chapter. Recommendations on the use of compulsory reading, optional reading and exercises are elaborated on in the preface of the first edition.

In general relativity, there are many verbose formulae, and the adoption of the system of geometrized units (where $c = 1$, $G = 1$) can greatly simplify these formulae. This system will also be used throughout this book. To help our readers understand the transition between geometric and non-geometric systems in a better way, we have attached an appendix (Appendix A).

I want to express my sincere thanks to Academician Ti-Pei Li, Academician Tan Lu, Researcher Han-Ying Guo, Prof. Liao Liu, Prof. Zheng Zhao, Researcher Runqiu Liu, Associate Professor Yongge Ma, Prof. Xuejun Yang and Prof. Guihua Tian, along with many other colleagues and readers for their contribution and support. I also want to thank readers of the first edition for their concern and love.

April 2005

Canbin Liang
Beijing Normal University
Beijing, China

¹ The second volume of the second edition was further divided into two volumes when it was published, which makes the entire work into three volumes.

Preface to the First Edition

Beginning in 1981, I was a visiting scholar at the relativity group of the University of Chicago for two years. Before going abroad, for various reasons, I only knew a little about general relativity, and even less about its essential mathematical tool, modern differential geometry. Thanks to the strong academic atmosphere of the relativity group of the University of Chicago, and thanks to the careful guidance of both Professor Robert Wald (my advisor) and Professor Robert Geroch, I soon became very interested in this research field. As a teacher, before I returned to China, I had a strong urge to teach my students what I had learned in the past two years as much as possible. As soon as I got back to China, I taught a series of graduate-level courses, starting with “Differential Geometry and General Relativity”. I was also invited to give lectures outside of Beijing. The past decade’s lecture notes have become the main source for writing this work. Over the past decade, I taught and learned to further deepen my understanding of the material I had been teaching. When confronted with difficulties, I would write to my mentors, Professor Wald and Professor Geroch, for help, and each time they gave me warm replies. Their insights would never fail to enlighten me. Physicists often feel that modern differential geometry is abstract and arcane at first pass, and fail to grasp its essence immediately. I think maybe I can help them with this issue. As I was once a first-time learner, I can empathize with how difficult it is to begin learning differential geometry. In addition, my past teaching experience may help reduce the subject’s difficulty. Reducing difficulty has not only become an aim of my past decade’s teaching, but also has become a major principle in the writing of this work. In order to reduce the difficulty, I spared no effort to elaborate, which dramatically increased this work’s length.

Modern differential geometry is not only crucial to the study of general relativity, but also has important applications in many sub-disciplines of physics (and even engineering). Many physicists have realized that modern differential geometry will play an increasingly important role in their further study based on results from international conferences and a substantial volume of literature, but find it difficult to learn the material properly. The heads of the Department of Physics of Beijing Normal University have recognized the importance of modern differential geometry to physicists much earlier. They encouraged and supported me to transfer my first graduate

course, Differential Geometry and General Relativity, to a one-semester elective course for advanced undergraduates (about 70 lecture hours) since 1995. More than half of the lecture hours are used for introducing basic knowledge of differential geometry (which corresponds to the first five chapters of this book). More than half of the remainder explains how to apply differential geometry to analyze special relativity (that is Chapter 6 of this book). The final part gives a brief introduction to general relativity (part of Chap. 7). Practice shows that physics undergraduates who like abstract thinking and have learned calculus and the basics of linear algebra can pass the final exam, as long as they attend class and spend time reviewing what they have learned and completing their homework assignments (about 5 questions per week on average). I am thrilled to find that some undergraduates (including sophomores) can understand the essence and develop a strong interest in the topics they learned. These undergraduates continued to take and excel in the graduate courses I offered, which covered all of the content following Sect. 7.4 of this volume.

This work is divided into two volumes. The first volume has 10 chapters. Among them, the first five chapters are an introduction to differential geometry. We begin to apply this knowledge in Chap. 6, in which we analyze special relativity. The last four chapters introduce the basic content of general relativity. Although the material and writing style of the first five chapters are geared toward those studying relativity, physicists who are not in the field of relativity can also use it as an introduction to differential geometry. The second volume further explores the advanced topics of general relativity (focusing on global analysis, such as the global causal structures of spacetime, asymptotically flat spacetimes, gravitational collapse, Kerr-Newman black holes, the $3 + 1$ decomposition of spacetimes, and the Lagrangian and Hamiltonian formulations of general relativity), as well as the mathematical tools required (such as conformal transformations, Lie groups and Lie algebras). The two volumes together can be used as a textbook for a graduate course and a reference for relativists. The first volume can also be used as a textbook for an advanced undergraduate elective course.

As to the writing style, another feature of this work is that it contains two parts—compulsory reading and optional reading—to meet the needs of readers at different levels. The compulsory reading is typed in SimSum while each optional reading is typed in KaiTi² and marked with the words [Optional Reading] and [The End of Optional Reading]. The compulsory reading is selfcontained, and skipping the optional readings does not affect the understanding of the subsequent compulsory reading. Footnotes may be treated similarly to optional readings. Beginners are advised to skip all the optional readings and footnotes during their first reading.

There are a large number of exercises in each chapter; however, their difficulty varies. In front of the title number, the most difficult exercises are marked with an asterisk *, which refers to the difficulty and does not mean that the problem is related to the optional reading. Questions marked with a tilde ~ before their title numbers are basic exercises that are strongly recommended. Among them, there are quite easy questions and some relatively difficult ones. In order to reduce the difficulty, hints

² In the English edition, each optional reading is indented and typed in a smaller font size.

are offered on those difficult questions. If time does not allow, the reader may choose to complete some of the questions that are marked with a tilde. It is okay to read the material without doing any exercises, but it is likely that you will find it difficult to understand the later chapters due to the lack of a strong foundation.

Owing to my limited knowledge and understanding, there may exist mistakes and deficiencies in this book. As an important way to reduce mistakes and deficiencies, I invited a large number of experts, colleagues and students to read part of the manuscript of the first volume. In alphabetical order by their last names, they are Bin Ao, Zhoujian Cao, Luru Dai, Xianxin Dai, Changjun Gao, Sijie Gao, Han He, Bo Hu, *Chao-Guang Huang, Zhiqian Kuang, **Liao Liu, Xiaoqin Li, Yongge Ma, Junjie Nan, **Shouyong Pei, **Wen-Chao Qiang, Hua Shen, *Qingjun Tian, *Xiaocen Tian, Bo-Bo Wang, Jinshan Wu, Xiaoning Wu, **Kongqing Yang, **Yun-Qiang Yu, *Xuejun Yang, Hongbao Zhang, Peng Zhang, Bin Zhou and Zong-Hong Zhu. (Those marked with ** are professors or researchers, and those marked with * are associate professors or associate researchers.) Those mentioned above have put forward many valuable suggestions on the chapters they read. I would like to express special thanks to two friends who helped me and contributed substantially to the writing of this work. The first one is Professor Robert Wald of the University of Chicago. He is an excellent advisor who enlightened me and provided me immense help with my teaching and writing after I returned to China. His masterpiece *General Relativity* is one of the major references for these volumes. The other one is Researcher Zhiqian Kuang from the Institute of Mathematics, CAS. He has reviewed many chapters of this book and put forward many important suggestions. Besides that, his profound thinking and deep understanding always benefited me when I discussed with him.

I would also like to express thanks to Professor Liao Liu from the Department of Physics of Beijing Normal University, and Professor Yuanxing Gui from the Department of Physics of Dalian University of Technology. Because of their recommendations, this work was included in the publishing plan of Beijing Normal University Press and received financial support from the press as well. I would like to sincerely thank Professor Zheng Zhao and Professor Yongcheng Wang for their care and support during the writing and publication. Besides that, I want to express thanks to Guifu Li, an editor of Beijing Normal University Press, for his support and contribution. I am grateful to the Beijing Municipal Education Commission for their approval and funding. I also appreciate the financial support provided by Beijing Normal University Press.

February 2000

Canbin Liang
Beijing Normal University
Beijing, China

Contents

1	Topological Spaces in Brief	1
1.1	The ABCs of Set Theory	1
1.2	Topological Spaces	6
1.3	Compactness [Optional Reading]	13
Exercises		16
Reference		17
2	Manifolds and Tensor Fields	19
2.1	Differentiable Manifolds	19
2.2	Tangent Vectors and Tangent Vector Fields	23
2.2.1	Tangent Vectors	23
2.2.2	Tangent Vector Fields on Manifolds	32
2.3	Dual Vector Fields	37
2.4	Tensor Fields	42
2.5	Metric Tensor Fields	47
2.6	The Abstract Index Notation	55
Exercises		62
References		65
3	The Riemann (Intrinsic) Curvature Tensor	67
3.1	Derivative Operators	67
3.2	Derivative and Parallel Transport of a Vector Field Along a Curve	75
3.2.1	Parallel Transport of a Vector Field Along a Curve	75
3.2.2	The Derivative Operator Associated with a Metric	78
3.2.3	Relationship Between the Derivative and Parallel Transport of a Vector Field Along a Curve	80
3.3	Geodesics	83
3.4	The Riemann Curvature Tensor	92
3.4.1	Definition and Properties of the Riemann Curvature	92
3.4.2	Computing Riemann Curvature from a Metric	97

3.5	The Intrinsic Curvature and the Extrinsic Curvature	99
	Exercises	101
	References	103
4	Lie Derivatives, Killing Fields and Hypersurfaces	105
4.1	Maps of Manifolds	105
4.2	Lie Derivatives	110
4.3	Killing Vector Fields	113
4.4	Hypersurfaces	119
	Exercises	126
	References	128
5	Differential Forms and Their Integrals	129
5.1	Differential Forms	129
5.2	Integration on Manifolds	134
5.3	Stokes's Theorem	138
5.4	Volume Elements	141
5.5	Integrating Functions on Manifolds, Gauss's Theorem	145
5.6	Dual Differential Forms	149
5.7	Computing the Riemann Curvature Using the Tetrad Method [Optional Reading]	153
	Exercises	160
	References	161
6	Special Relativity	163
6.1	Foundations of the 4-Dimensional Formulation	163
6.1.1	Preliminaries	163
6.1.2	The Background Spacetime of Special Relativity	165
6.1.3	Inertial Observers and Inertial Frames	167
6.1.4	Proper Time and Coordinate Time	170
6.1.5	Spacetime Diagrams	172
6.1.6	Spacetime Structure: Special Relativity Versus Pre-Relativity Physics	175
6.2	Interesting Typical Effects	179
6.2.1	Length Contraction	179
6.2.2	Time Dilation	181
6.2.3	The Twin "Paradox"	186
6.2.4	The Garage "Paradox"	188
6.3	Kinematics and Dynamics of a Point Mass	189
6.4	The Energy-Momentum Tensor of Continuous Media	206
6.5	Perfect Fluid Dynamics	211
6.6	Electrodynamics	216
6.6.1	Electromagnetic Fields and 4-Current Densities	216
6.6.2	Maxwell's Equations	220
6.6.3	Lorentz 4-Force	223

6.6.4	The Energy-Momentum Tensor of an Electromagnetic Field	225
6.6.5	Electromagnetic 4-Potential and Its Equation of Motion, Electromagnetic Waves	226
6.6.6	The Doppler Effect on a Light Wave	233
Exercises		234
References		237
7	Foundations of General Relativity	239
7.1	Gravity and Spacetime Geometry	239
7.2	Physical Laws in Curved Spacetime	244
7.3	Fermi-Walker Transport and Non-Rotating Observers	249
7.4	The Proper Coordinate System of an Arbitrary Observer	259
7.5	Equivalence Principles and Local Inertial Frames	267
7.6	Tidal Forces and the Geodesic Deviation Equation	273
7.7	The Einstein Field Equation	282
7.8	Linear Approximation and the Newtonian Limit	287
7.8.1	Linearized Theory of Gravity	287
7.8.2	The Newtonian Limit	292
7.9	Gravitational Radiation	296
7.9.1	Gauge Conditions of the Linearized Theory of Gravity	296
7.9.2	Gravitational Plane Waves	302
7.9.3	Emission of Gravitational Waves	316
7.9.4	Detection of Gravitational Waves	317
Exercises		326
References		328
8	Solving Einstein's Equation	331
8.1	Stationary Spacetimes and Static Spacetimes	331
8.2	Spherically Symmetric Spacetimes	336
8.3	The Vacuum Schwarzschild Solution	340
8.3.1	Static Spherically Symmetric Metrics	340
8.3.2	The Vacuum Schwarzschild Solution	341
8.3.3	Birkhoff's Theorem	347
8.4	The Reissner-Nordström Solution	349
8.4.1	Electrovacuum Spacetimes and the Einstein-Maxwell Equations	349
8.4.2	The Reissner-Nordström Solution	351
8.5	Axisymmetric Metrics [Optional Reading]	355
8.6	Plane Symmetric Metrics [Optional Reading]	357
8.7	The Newman-Penrose (NP) Formalism [Optional Reading]	360
8.8	Solving the Einstein-Maxwell Equations Using the NP Formalism [Optional Reading]	367
8.8.1	Maxwell's Equations and Einstein's Equation in the NP Formalism	367

8.8.2	An Example of Solving the Einstein-Maxwell Equations Under the Axisymmetric Condition	370
8.9	The Vaidya Metric and the Kinnersley Metric	378
8.9.1	From the Schwarzschild Metric to the Vaidya Metric	378
8.9.2	The Kinnersley Metric	383
8.9.3	The Kinnersley Metric (Detailed Discussions)	387
8.10	Coordinate Conditions, the Gauge Freedom of General Relativity	396
8.10.1	Coordinate Conditions	396
8.10.2	The Gauge Freedom of General Relativity	401
	Exercises	404
	References	405
9	Schwarzschild Spacetimes	407
9.1	Geodesics in Schwarzschild Spacetimes	407
9.2	Classical Experimental Tests of General Relativity	412
9.2.1	Gravitational Redshift	413
9.2.2	Perihelion Precession of Mercury	415
9.2.3	Light Deflection	419
9.3	Spherical Stars and Their Evolution	422
9.3.1	Interior Solutions for Static Spherical Stars	422
9.3.2	Stellar Evolution	430
9.4	The Kruskal Extension and Schwarzschild Black Holes	439
9.4.1	The Definition of a Spacetime Singularity	440
9.4.2	Coordinate Singularities of Rindler Metrics	442
9.4.3	The Kruskal Extension of Schwarzschild Spacetimes	446
9.4.4	Surfaces of Infinite Redshift in Schwarzschild Spacetimes	453
9.4.5	Embedding Diagrams [Optional Reading]	454
9.4.6	The Gravitational Collapse of a Spherical Star and Schwarzschild Black Holes	456
	Exercises	463
	References	465
10	Cosmology I	467
10.1	Kinematics of the Universe	468
10.1.1	Cosmological Principle	468
10.1.2	Spacial Geometries of the Universe	470
10.1.3	The Robertson-Walker Metric	480
10.2	Dynamics of the Universe	488
10.2.1	The Hubble-Lemaître Law	488
10.2.2	Cosmological Redshift	490

Contents	xxiii
10.2.3 Evolution of the Scale Factor	493
10.2.4 The Cosmological Constant and Einstein's Static Universe	501
10.3 The Thermal History of Our Universe	503
10.3.1 A Brief History of the Universe	503
10.3.2 The Dark Matter Problem	516
10.3.3 The Cosmological Constant Problem and the Λ CDM Model	520
Exercises	523
References	524
Appendix A: The Conversion Between Geometrized and Nongeometrized Unit Systems	527
Conventions and Notation	535
Index	539

Outline of Volume II

11 Global Causal Structure of Spacetime

- 11.1 Pasts and Futures
- 11.2 Inextendible Causal Curves
- 11.3 Causality Conditions
- 11.4 Domains of Dependence
- 11.5 Cauchy Surfaces, Cauchy Horizons and Globally Hyperbolic Spacetimes

Exercises

12 Asymptotically Flat Spacetimes

- 12.1 Conformal Transformations
- 12.2 The Conformal Infinity of Minkowski Spacetime
- 12.3 The Conformal Infinity of Schwarzschild Spacetimes
- 12.4 Isolated systems and Asymptotically Flat Spacetimes
- 12.5 Symmetries on \mathcal{I}^\pm and i^0 , the BMS Group and SPI Group
- 12.6 The Non-locality of Gravitational Energy
- 12.7 The Total Energy and Total Momentum of an Asymptotically Flat Spacetime

Exercises

13 Kerr-Newman (KN) Black Holes

- 13.1 Reissner-Nordström (RN) Black Holes
- 13.2 The Kerr-Newman Metric
- 13.3 The Maximum Extension of KN Spacetimes
- 13.4 Static Limit Surfaces, Ergospheres and More
- 13.5 Extracting Energy from a Rotating Black Hole (Penrose Process)
- 13.6 The Black Hole “No-Hair” Conjecture

Exercises

14 Revisiting Reference Frames

- 14.1 General Discussions on Reference Frames
- 14.2 Einstein’s Rotating Disk
- 14.3 Clock Synchronization in a Reference Frame [Optional Reading]
- 14.4 The 3 + 1 Decomposition of Spacetimes
- 14.5 An Application of the 3 + 1 Decomposition of Spacetimes—the Initial Value Problem in General Relativity

Exercises

15 Cosmology II

- 15.1 Finding a Way Out of the Difficulties of the Standard Cosmological Model—the Inflationary Model
- 15.2 Dark Matter
- 15.3 The Cosmological Constant and the Vacuum Energy Problem
- 15.4 Dark Energy and the “New Standard Cosmological Model”

Exercises

Appendix B Mathematical Foundation of Quantum Mechanics in Brief

- B.1 The ABCs of a Hilbert Space
- B.2 Unbounded Operators and Their Self-Adjointness

Exercises

Appendix C Geometric Phases in Quantum Mechanics

- C.1 Berry’s Geometric Phase
- C.2 The Aharonov-Anandan (AA) Geometric Phase

Appendix D Energy Conditions**Appendix E Singularity Theorems and the Cosmic Censorship Hypotheses**

- E.1 Introducing the Singularity Theorems
- E.2 Cosmic Censorship Hypotheses
- E.3 The Strong Cosmic Censorship Hypothesis in Terms of TIPs [Optional Reading]
- E.4 Singular Boundaries

Appendix F The Frobenius Theorem**Appendix G Lie Groups and Lie Algebras**

- G.1 The ABCs of Group Theory
- G.2 Lie Groups
- G.3 Lie Algebras
- G.4 One Parameter Subgroups and Exponential Maps
- G.5 Important Lie Groups and Lie Algebras

- G.6 Structure Constants of a Lie Algebra
 - G.7 Lie Groups of Transformations and Killing Vector Fields
 - G.8 Adjoint Representations and Killing Forms [Optional Reading]
 - G.9 The Proper Lorentz Group and the Lorentz Algebra
- Exercises

Outline of Volume III

16 Lagrangian and Hamiltonian Formulations of General Relativity

- 16.1 Lagrangian Formalism
 - 16.2 Hamiltonian Formalism for Systems with a Finite Number of Degrees of Freedom
 - 16.3 Expressing Lagrangian and Hamiltonian Formalisms of Systems with a Finite Number of Degrees of Freedom in Geometric Language [Optional Reading]
 - 16.4 Hamiltonian Formulation of Classical Field Theories
 - 16.5 Hamiltonian Formulation of General Relativity
 - 16.6 Tensor Densities [Optional Reading]
 - 16.7 Symplectic Geometry and Its Applications in the Hamiltonian Formalism [Optional Reading]
 - 16.8 From Geometrodynamics to Connection Dynamics—A Brief Introduction to Ashtekar’s New Variables [Optional Reading]
- Exercises

17 Isolated Horizons, Dynamical Horizons and Black Hole Thermodynamics

- 17.1 Traditional Black Hole Thermodynamics and Its Shortcomings
 - 17.2 Null Geodesic Congruences and Their Raychaudhuri Equations
 - 17.3 The Raychaudhuri Equations on a Null Hypersurface
 - 17.4 Trapped Surfaces and Apparent Horizons
 - 17.5 Weakly Isolated Horizons and Their Zeroth and First Laws
 - 17.6 Further Discussions on Weakly Isolated Horizons [Optional Reading]
 - 17.7 Dynamical Horizons and Their Mechanical Laws
- Exercises

**Appendix H Spacetime Symmetries and Conservation Laws
(Noether's Theorem)**

- H.1 Proving Noether's Theorem Using Geometric Language
- H.2 Canonical Energy-Momentum Tensors
- H.3 On the Proof Using Coordinate Language

Appendix I Fiber Bundles and Their Applications in Gauge Theories

- I.1 Principal Fiber Bundles
- I.2 Connections on a Principal Fiber Bundle
- I.3 Fiber Bundles Associated to a Principal Fiber Bundle
(Associated Bundles)
- I.4 The Global Gauge Invariance of Physical Fields
- I.5 The Local Gauge Invariance of Physical Fields
- I.6 The Physical Meaning of Cross Sections
- I.7 Gauge Potential and Connection
- I.8 Gauge Field Strength and Curvature
- I.9 Connections and Covariant Derivatives on a Vector Bundle

Exercises

Appendix J De Sitter Spacetime and Anti-de Sitter Spacetime

- J.1 Spaces of Constant Curvature
- J.2 De Sitter Spacetime
- J.3 The Penrose Diagram of de Sitter Spacetime
- J.4 More on Event Horizons and Particle Horizons
- J.5 Schwarzschild-de Sitter Spacetime
- J.6 Anti-de Sitter Spacetime

Chapter 1

Topological Spaces in Brief



1.1 The ABCs of Set Theory

A well-determined collection of some amount of objects is called a set. Each object in the set is called an **element** or a **point**. If x is an element of a set X , then we say “ x belongs to X ”, and denote it by $x \in X$. The symbol \notin stands for “does not belong to”. There are two ways to express a set, one is to list each of its elements, separated by commas, and enclosing all elements in a curly brackets; for example,

$$X = \{1, 4, 5.6\}$$

represents the set that consists of the real numbers 1, 4 and 5.6. The other expression is to point out the general character of elements in a set; for example,

$$X = \{x \mid x \text{ is a real number}\}$$

represents the set of all real numbers (the common notation of this specific set is \mathbb{R}), while

$$X = \{x \in \mathbb{R} \mid x > 9\}$$

represents the set of all real numbers that are greater than 9.

The set that has no elements is called the **empty set**, denoted by \emptyset .

Definition 1 If each element of a set A belongs to a set X , then we say A is a **subset** of X . We also say that A is contained in X , or X contains A , denoted by $A \subset X$ or $X \supset A$. Stipulate that \emptyset is a subset of any set. Two sets X and Y are said to be **equal** (denoted by $X = Y$) if $X \subset Y$ and $Y \subset X$. A is called a **proper subset** of X (denoted by $A \subsetneq X$) if $A \subset X$ and $A \neq X$.

Remark 1 A more exact expression of the definition of a subset should be “a set S is called a subset of the set X if and only if each element of A belongs to X ”. However,

for convenience's sake, all of the terms "if" and "when" in the definition are implied to indicate "if and only if".

This text will use \coloneqq to represent "is defined as" and use \equiv to represent "identical to" or "denoted by". For example, $C \equiv A - B$ means "denote $A - B$ by C ". The adoption of these two symbols is simply for clarity, they may be replaced by the equal sign as well.

Definition 2 The union, intersection, difference and complement of two sets A and B are defined as follows:

Union $A \cup B := \{x \mid x \in A \text{ or } x \in B\}$.

Intersection $A \cap B := \{x \mid x \in A, x \in B\}$. (The condition " $x \in A, x \in B$ " is short for " $x \in A$ and $x \in B$ ", the same applies below.)

Difference $A - B := \{x \mid x \in A, x \notin B\}$. (Mathematics books usually denote the difference by $A \setminus B$ or $A \sim B$, this text will denote all of them by $A - B$.)

If A is a subset of X then $-A$, the **complement** of A , is defined as $-A := X - A$.

Theorem 1.1.1 *The operations above obey the following laws:*

Commutative law $A \cup B = B \cup A, A \cap B = B \cap A$.

Associative law $(A \cup B) \cup C = A \cup (B \cup C), (A \cap B) \cap C = A \cap (B \cap C)$.

Distributive law $(A \cap B) \cup C = (A \cup C) \cap (B \cup C), (A \cup B) \cap C = (A \cap C) \cup (B \cap C)$.

De Morgan's law $A - (B \cup C) = (A - B) \cap (A - C), A - (B \cap C) = (A - B) \cup (A - C)$.

Proof As an example, we prove the second equation of De Morgan's Law (the reader should verify the rest). All we have to do for this is to show that both sides of the equation contain one another.

(A) Suppose $x \in A - (B \cap C)$, then $x \in A, x \notin B \cap C$. The latter leads to $x \notin B$ or $x \notin C$. Combining $x \in A$ and $x \notin B$, we have $x \in A - B$; combining $x \in A$ and $x \notin C$, we have $x \in A - C$, and hence $x \in (A - B) \cup (A - C)$. Therefore,

$$A - (B \cap C) \subset (A - B) \cup (A - C).$$

(B) Suppose $x \in (A - B) \cup (A - C)$, then $x \in A - B$ or $x \in A - C$. The former leads to $x \in A, x \notin B$; the latter leads to $x \in A, x \notin C$. Combining the two, we have $x \in A, x \notin B \cap C$, and hence $x \in A - (B \cap C)$. Therefore,

$$(A - B) \cup (A - C) \subset A - (B \cap C).$$

□

Definition 3 The **Cartesian product** $X \times Y$ of two non-empty sets X and Y is defined as

$$X \times Y := \{(x, y) \mid x \in X, y \in Y\}.$$

That is, $X \times Y$ is a set such that each of its element is an ordered pair (x, y) formed by one element x from X and one element y from Y . The Cartesian product of a

(finite¹) number of sets can be defined similarly; for example,

$$X \times Y \times Z := \{(x, y, z) \mid x \in X, y \in Y, z \in Z\}.$$

We also stipulate that the Cartesian product satisfies the associative law, i.e., $(X \times Y) \times Z = X \times (Y \times Z)$.

Example 1 $\mathbb{R}^2 := \mathbb{R} \times \mathbb{R}$, $\mathbb{R}^n := \mathbb{R} \times \cdots \times \mathbb{R}$ (n factors in total). Since an element of \mathbb{R}^2 is an ordered pair formed by two real numbers, these two real numbers are called the **natural coordinates** of this element. Similarly, each element of \mathbb{R}^n has n natural coordinates. It follows that \mathbb{R}^n is intrinsically endowed with coordinates, though this is not necessarily true for other sets. Using natural coordinates, the concept of distance between any two elements of \mathbb{R}^n can be defined.

Definition 4 The **distance**, denoted by $|y - x|$, between any two elements $x = (x^1, \dots, x^n)$ and $y = (y^1, \dots, y^n)$ is defined as

$$|y - x| := \sqrt{\sum_{i=1}^n (y^i - x^i)^2}.$$

Starting from the next paragraph, this text will use the mathematical symbols \forall (stands for “for all” or “for any”) and \exists (stands for “there exists”) frequently, please be familiar with them.

Definition 5 Suppose X and Y are non-empty sets. A **map** from X to Y (denoted by $f : X \rightarrow Y$) is a rule that associates each element of X with a unique element in Y . If $y \in Y$ is the corresponding element of $x \in X$, we write $y = f(x)$; we also call y the **image** of x under the map f and call x the **preimage** (or **inverse image**) of y . X is called the **domain** of the map f , and the set of images of all elements of X (denoted by $f[X]$) is called the **range** of the map $f : X \rightarrow Y$. Maps $f : X \rightarrow Y$ and $f' : X \rightarrow Y$ are said to be **equal** if $f(x) = f'(x) \forall x \in X$.

Remark 2 Usually, $y = f(x)$ is also written as $f : x \mapsto y$. Please note the difference between \mapsto and \rightarrow : \rightarrow in $f : X \rightarrow Y$ means that f is a map from X to Y (a set to a set), while \mapsto in $f : x \mapsto y$ means that the image of $x \in X$ under the map f is y (an element to an element).

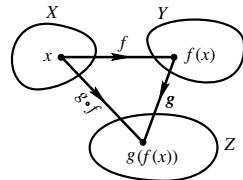
Remark 3 Suppose $A \subset X$, then the subset formed by the images of elements of A under the map f is denoted by $f[A]$, i.e.,

$$f[A] \equiv \{y \in Y \mid \exists x \in A \text{ such that } y = f(x)\} \subset Y.$$

Example 2 A single-valued function $y = f(x)$ in standard calculus is a map from \mathbb{R} (or its subset) to \mathbb{R} .

¹ The Cartesian product of an infinite number of sets can also be defined, but this is outside the scope of this text.

Fig. 1.1 Composite map $g \circ f$. NB: First perform f then perform g



Remark 4 A map from \mathbb{R}^2 to \mathbb{R} gives a function of two variables, because each point in \mathbb{R}^2 is described by two real numbers (natural coordinates). Similarly, a map from \mathbb{R}^n to \mathbb{R}^m gives m functions of n variables.

Definition 6 A map $f : X \rightarrow Y$ is said to be **one-to-one** if there is no more than one inverse image for any $y \in Y$ (there may be none). $f : X \rightarrow Y$ is said to be **onto** if there is at least one inverse image for any $y \in Y$ (there may be more than one).²

Remark 5 ① A necessary and sufficient condition for f to be onto is that the range $f[X] = Y$. ② If f is a one-to-one map, then there exists an inverse map $f^{-1} : f[X] \rightarrow X$. However, whether f has an inverse map or not, we can always define $f^{-1}[B]$, the “inverse image” of any subset $B \subset Y$ under f , as

$$f^{-1}[B] := \{x \in X \mid f(x) \in B\} \subset X .$$

Note that the “inverse” here is a subset (rather than an element) of X . For example, if X has (and only has) two elements x and x' , whose image under the action of f are both $y \in Y$, then although the inverse map $f^{-1} : Y \rightarrow X$ does not exist, $f^{-1}[\{y\}]$ is still meaningful when considering y as a one-point subset (i.e., $\{y\}$) of Y , and the meaning is $f^{-1}[y] = \{x, x'\} \subset X$.

Definition 7 $f : X \rightarrow Y$ is called a **constant map**, if $f(x) = f(x') \ \forall x, x' \in X$.

Definition 8 Suppose X, Y, Z are sets and that $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ are maps, then the **composite map** denoted by $g \circ f$ is a map from X to Z , which is defined as $(g \circ f)(x) := g(f(x)) \in Z \ \forall x \in X$ (see Fig. 1.1).

Remark 6 If $X = Y = Z = \mathbb{R}$, then the composite map $g \circ f$ is the familiar composite function of one variable.

If X and Y are just two sets (with no additional structure), “one-to-one” and “onto” are the only two requirements that we can impose on a map between X and Y ; however, if some kinds of additional structures are assigned to X and Y , then sometimes we can impose additional requirements on $f : X \rightarrow Y$. For example, we

² Many mathematics books call the one-to-one and onto maps used in this text **injections** and **surjections**, respectively, and call maps that are both injective and surjective one-to-one maps (also called **bijections**). Thus, their one-to-one maps are stronger than the one-to-one maps in this text.

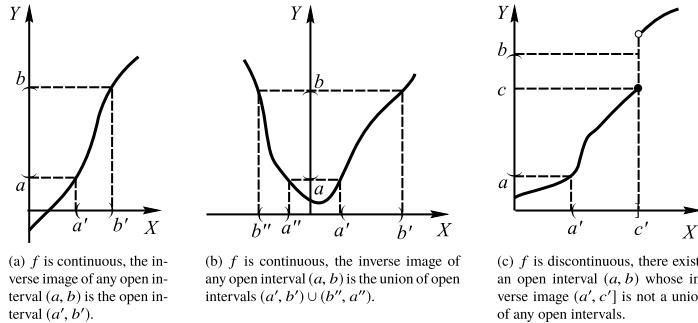


Fig. 1.2 Expressing continuity in terms of open intervals, the curve represents the map $f : X \rightarrow Y$

can ask $f : \mathbb{R} \rightarrow \mathbb{R}$ to be continuous, or even smooth. The continuity of a function of one variable is already defined in calculus (the “ $\varepsilon - \delta$ definition”), we restate it as follows: ① f is said to be continuous at x if $\forall \varepsilon > 0 \exists \delta > 0$ such that $|f(x') - f(x)| < \varepsilon$ when $|x' - x| < \delta$; ② f is said to be continuous on \mathbb{R} if it is continuous at any point of \mathbb{R} . It seems that this definition is unable to be generalized to a map between two sets that have no definition of distance, since it depends on the concept of distance between any two elements in \mathbb{R} (for \mathbb{R} , the distance is the difference of the natural coordinates). However, after pondering this, one can find that the $\varepsilon - \delta$ definition can be rephrased with the concept of open intervals (rather than the concept of distance) as follows: suppose $X = Y = \mathbb{R}$, the map $f : X \rightarrow Y$ is said to be continuous if the “inverse image” of any open interval in Y is a union of open intervals in X (or an empty set). The equivalence between this statement and the familiar $\varepsilon - \delta$ statement may be illustrated by Fig. 1.2 (we do not mean to prove it here). In Fig. 1.2a, the map $f : X \rightarrow Y$ is continuous according to the $\varepsilon - \delta$ definition; correspondingly, the inverse image of an arbitrary open interval (a, b) in Y is the open interval (a', b') . In Fig. 1.2b, the map is continuous; correspondingly, the inverse image of an arbitrary open interval (a, b) in Y is the union of open intervals (a', b') and (b'', a'') . In Fig. 1.2c, the map $f : X \rightarrow Y$ is discontinuous at $c' \in X$; correspondingly, there exists an open interval (a, b) in Y whose inverse image $f^{-1}[(a, b)] = (a', c') \subset X$ is neither an open interval nor a union of open intervals. The discussion above indicates one aspect of the use of the concept “the union of open intervals”: to define the continuity of the map $f : \mathbb{R} \rightarrow \mathbb{R}$. Actually this concept also has many other uses, so it is often necessary to generalize it to any set X other than \mathbb{R} . For convenience, we refer to any subset of \mathbb{R} that can be expressed as the union of open intervals (along with the empty set \emptyset) an open subset. To generalize the concept of an open subset to any set X , we should first determine the essential, abstract (thus generalizable) properties of the open subsets of \mathbb{R} . They are: (a) both \mathbb{R} and \emptyset are open subsets; (b) the intersection of a finite number of open subsets is still an open subset; (c) the union of any number of open subsets is still an open subset. After generalizing these three properties, we can define the concept of an open subset for any set X . A set with open subsets defined is called a topological space. From the concept of open subsets,

one can also define many concepts and prove many theorems, which develop into a complete and fruitful branch of mathematics—point-set topology. Sections 1.2 and 1.3 will give an introduction to the basics of topological spaces.

1.2 Topological Spaces

As we mentioned at the end of Sect. 1.1, subsets of \mathbb{R} can be divided into two categories: open subsets and non-open subsets. (Any subset is either open or non-open. Do not refer to a non-open subset as a closed subset. According to the definition of a closed subset that we will introduce later, a subset can be open, closed, both, or neither.) The collection of open subsets has the above three properties (a), (b), (c). For any non-empty set X , we can also assign some subsets to be open and others to be non-open in an appropriate manner. To make this assignment useful, we require that any method of assignment should satisfy three requirements: (a) X itself and the empty set \emptyset are open subsets; (b) the intersection of a finite number of open subsets is an open subset; (c) the union of any number (which may be finite or infinite) of open subsets is an open subset. For a given set, there are usually many ways of assigning openness that meet these three requirements. For example, suppose X is a set, we can assign X and \emptyset as open subsets, and all others as non-open. This certainly satisfies the above three requirements, with the feature that it has the lowest number of open subsets (only two). However, we can also have another extreme assignment, namely to assign any subset of X to be an open subset. It is not hard to see that this method of assigning openness also satisfies the three requirements above. Although the two assignments above do not necessarily have much use, they at least can indicate that there is more than one way of assigning openness to meet the three requirements above. We say that each of the assignments which satisfies the three requirements above gives an additional structure to the set X , called the topological structure. For a set with a topological structure defined, we can point at any subset of it and ask: “Is this an open subset?” The answer will be either “yes” or “no”, with no middle ground. Conversely, for a set without any topological structure defined, this kind of question is meaningless. If X is a set having a topological structure, then the collection of its open subsets also form a set, called a topology on X , denoted by \mathcal{T} . Let \mathcal{P} represent the collection of all subsets of X (as shown in Fig. 1.3), then any open subset O and any non-open subset V are both elements in \mathcal{P} . All of the open subsets of X form a subset \mathcal{T} of \mathcal{P} (note that it is not a subset of X), which is the topology on X . Please notice the difference between the symbol \subset and \in : $O \subset X$ only indicates that O is a subset of X , while $O \in \mathcal{T}$ indicates that O is an open subset of X . The discussion above will pave the way for understanding the definitions expressed by the following mathematical language.

Definition 1 A topology \mathcal{T} on a non-empty set X is a collection of some subsets of X , which satisfies:

- (a) $X, \emptyset \in \mathcal{T}$;

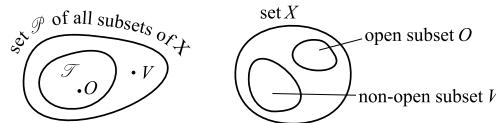


Fig. 1.3 \mathcal{P} is the collection of all subsets of X . Any subset of X (e.g., O, V) is an element of \mathcal{P} . \mathcal{T} is a subset of \mathcal{P} such that each element of it (e.g., O) is an open subset of X

(b) If $O_i \in \mathcal{T}, i = 1, 2, \dots, n$, then $\bigcap_{i=1}^n O_i \in \mathcal{T}$ (where $\bigcap_{i=1}^n O_i$ stands for the intersection of these O_i);

(c) If $O_\alpha \in \mathcal{T} \forall \alpha$, then $\bigcup_\alpha O_\alpha \in \mathcal{T}$. (Adding $\forall \alpha$ after $O_\alpha \in \mathcal{T}$ indicates that each O_α belongs to \mathcal{T} , with no restriction on the number of O_α . $\bigcup_\alpha O_\alpha \in \mathcal{T}$ indicates that the union of all O_α belongs to \mathcal{T} .)

Definition 2 A set X with a topology \mathcal{T} assigned to it is called a **topological space**. A subset O of the topological space X is called an **open subset** (or **open set** for short) if $O \in \mathcal{T}$.

For the same set X we can define different topologies (there may be many \mathcal{T} that satisfy Definition 1). Suppose \mathcal{T}_1 and \mathcal{T}_2 are both topologies on X , then a subset A of X may satisfy both $A \in \mathcal{T}_1$ and $A \notin \mathcal{T}_2$; that is, A is an open set for \mathcal{T}_1 (measured by \mathcal{T}_1), but not an open set for \mathcal{T}_2 . We thus see that \mathcal{T}_1 and \mathcal{T}_2 define X as two different topological spaces. In order to clarify the choice of the topology, we use (X, \mathcal{T}) to represent a topological space. As a result, (X, \mathcal{T}_1) and (X, \mathcal{T}_2) represent two different topological spaces, even though both of their “base sets” is X . After a topology is specified, one can also just use X to represent a topological space.

Which topology should we choose for a given set X to make it a topological space? This depends on the properties of X itself, as well as what kind of problem we are considering. For example, we may choose the so-called “usual topology” as the topology for the set \mathbb{R} in most of the problems we are usually concerned with (see Example 3 below).

Example 1 Suppose X is an arbitrary non-empty set, and let \mathcal{T} to be the collection of all subsets of X , then it obviously satisfies the three requirements in Definition 1, and hence forms a topology on X , which is called the **discrete topology**.

Example 2 Suppose X is an arbitrary non-empty set, and let $\mathcal{T} = \{X, \emptyset\}$, then it obviously satisfies the three requirements in Definition 1 and hence forms a topology on X , which is called the **indiscrete topology**. The indiscrete topology is the topology that has the lowest number of elements.

Example 3 (1) Suppose $X = \mathbb{R}$, then $\mathcal{T}_u := \{\text{the empty set or subsets of } \mathbb{R} \text{ that can be expressed as a union of open intervals}\}$ is called the **usual topology** on \mathbb{R} .

(2) Suppose $X = \mathbb{R}^n$, then $\mathcal{T}_u := \{\text{the empty set or subsets of } \mathbb{R}^n \text{ that can be expressed as a union of open balls}\}$ is called the **usual topology** on \mathbb{R}^n , where an **open ball** is defined as $B(x_0, r) := \{x \in \mathbb{R}^n \mid |x - x_0| < r\}$, x_0 is called the center

and $r > 0$ is called the radius. An open ball in \mathbb{R}^2 is also called an **open disk**; an open ball in \mathbb{R} is just an open interval.

It is not difficult to check that the \mathcal{T}_u in (1) and (2) satisfy the three requirements in Definition 1. According to the definition above, any open interval of \mathbb{R} is an open set measured by \mathcal{T}_u . However, in principle we can also choose other topologies to make \mathbb{R} a topological space different from $(\mathbb{R}, \mathcal{T}_u)$. For example, if measured by the indiscrete topology, then no subset is an open set other than \mathbb{R} and \emptyset . In contrast, if measured by the discrete topology, then any subset of \mathbb{R} (including any closed interval or half-closed interval) is an open set. From now on, we will consider $(\mathbb{R}^n, \mathcal{T}_u)$ when we treat \mathbb{R}^n as a topological space unless stated otherwise.

Example 4 Suppose (X_1, \mathcal{T}_1) and (X_2, \mathcal{T}_2) are topological spaces, $X = X_1 \times X_2$ (i.e., X is the Cartesian product of X_1 and X_2). Define the topology on X as

$$\begin{aligned}\mathcal{T} := \{O \subset X \mid O \text{ can be expressed in the form of a union of subsets of } O_1 \times O_2, \\ O_1 \in \mathcal{T}_1, O_2 \in \mathcal{T}_2\},\end{aligned}\tag{1.2.1}$$

then \mathcal{T} is called the **product topology** on X .

[Optional Reading 1.2.1]

It is not difficult to generalize the definition of the product topology from two topological spaces to a finite number of topological spaces. However, there is a point we should make: suppose $X = X_1 \times X_2 \times X_3$, then X can be regarded as either $(X_1 \times X_2) \times X_3$ or $X_1 \times (X_2 \times X_3)$. If $X_{12} = X_1 \times X_2$, and we use \mathcal{T}_{12} to represent its product topology, then we can define the product topology for the set $X = X_{12} \times X_3$, denoted by \mathcal{T} . Similarly, if $X_{23} = X_2 \times X_3$, and we use \mathcal{T}_{23} to represent its product topology, then we can also define the product topology for the set $X = X_1 \times X_{23}$, denoted by \mathcal{T}' . Following (1.2.1), we can also define the product topology $\tilde{\mathcal{T}}$ for $X = X_1 \times X_2 \times X_3$ as follows:

$$\tilde{\mathcal{T}} := \{O \subset X \mid O \text{ can be expressed in the form of a union of subsets of } O_1 \times O_2 \times O_3, \quad O_1 \in \mathcal{T}_1, O_2 \in \mathcal{T}_2, O_3 \in \mathcal{T}_3\}.$$

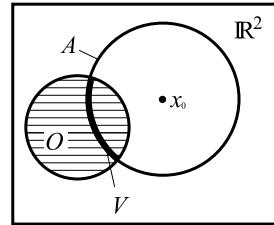
It can be proved that $\tilde{\mathcal{T}} = \mathcal{T}' = \mathcal{T}$; that is, these three definitions for the product topology on $X = X_1 \times X_2 \times X_3$ are different routes to the same end, so there is no ambiguity. The conclusion for the case of a finite number of topological spaces is also similar. A simple example is $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$, where one can also prove that the product topology defined in this way agrees with the topology defined in Example 3 in terms of open balls.

[The End of Optional Reading 1.2.1]

Example 5 Suppose (X, \mathcal{T}) is a topological space, and A is an arbitrary non-empty subset of X . If A is regarded as a set, we can certainly assign a topology, denoted by \mathcal{S} (script S), and thereby make it a topological space, denoted by (A, \mathcal{S}) . Since A is a subset of X , we would hope that \mathcal{S} and \mathcal{T} are closely related. If $A \in \mathcal{T}$, then the problem is simple, we just need to define $\mathcal{S} := \{V \subset A \mid V \in \mathcal{T}\}$. However, if $A \notin \mathcal{T}$, we have $A \notin \mathcal{S}$ according to the definition above, which contradicts condition (a) of Definition 1. Therefore, the definition of \mathcal{S} above is illegal. A smart definition is

$$\mathcal{S} := \{V \subset A \mid \exists O \in \mathcal{T} \text{ such that } V = A \cap O\}.\tag{1.2.2}$$

Fig. 1.4 The bold line segment (excluding the endpoints) is a subset V of A , since it can be considered as an intersection of $O \in \mathcal{T}_u$ and A . From (1.2.2) we can see that $V \in \mathcal{S}$



It can be proved from the equation above that $A \in \mathcal{S}$ even if $A \notin \mathcal{T}$, furthermore \mathcal{S} satisfies the other conditions of Definition 1 (see Exercise 1.6). The \mathcal{S} defined in this way is called the **induced topology** on $A(\subset X)$ derived from \mathcal{T} . Later on, unless stated otherwise, when we treat a subset A of (X, \mathcal{T}) as a topological space, we will consider it as (A, \mathcal{S}) , where \mathcal{S} is the topology induced by \mathcal{T} . (A, \mathcal{S}) is called a **topological subspace** of (X, \mathcal{T}) .

The example below is helpful for a better understanding of induced topology. Consider a unit circle S^1 in \mathbb{R}^2 defined by $S^1 := \{x \in \mathbb{R}^2 \mid |x - x_0| = 1\}$, whose center is at x_0 . Suppose $A \subset \mathbb{R}^2$ is S^1 , then A is not open measured by \mathcal{T}_u on \mathbb{R}^2 , since it cannot be expressed as an union of open balls in \mathbb{R}^2 (a line is too thin to fill in any open disk). If we define an induced topology \mathcal{S} for A using (1.2.2), then A is open as measured by \mathcal{S} . Moreover, suppose V is an arbitrary segment of A (excluding the two endpoints); then, as shown by the bold line in Fig. 1.4, although V is not an open set measured by \mathcal{T}_u , it is open as measured by \mathcal{S} , since there exists an open disk $O \in \mathcal{T}_u$ such that $V = A \cap O$.

Using the concept of open sets, we can define the continuity of maps between topological spaces. Two equivalent definitions are given below; the proof of the equivalence is left as an exercise (Exercise 1.10).

Definition 3a Suppose (X, \mathcal{T}) and (Y, \mathcal{S}) are topological spaces. A map $f : X \rightarrow Y$ is said to be **continuous** if $f^{-1}[O] \in \mathcal{T} \forall O \in \mathcal{S}$ (for the definition of $f^{-1}[O]$, see ② of Remark 5 in Sect. 1.1).

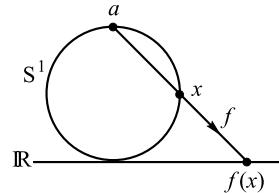
Definition 3b Suppose (X, \mathcal{T}) and (Y, \mathcal{S}) are topological spaces. A map $f : X \rightarrow Y$ is said to be **continuous at a point** $x \in X$ if $\forall G' \in \mathcal{S}$ that satisfies $f(x) \in G'$, $\exists G \in \mathcal{T}$ such that $x \in G$ and $f[G] \subset G'$. $f : X \rightarrow Y$ is said to be **continuous** if it is continuous at every point $x \in X$.

Remark 1 It is easy to see that if $X = Y = \mathbb{R}$, and $\mathcal{T} = \mathcal{S} = \mathcal{T}_u$, then Definition 3b (and thus Definition 3a) will return to the $\varepsilon - \delta$ definition.

Definition 4 Topological spaces (X, \mathcal{T}) and (Y, \mathcal{S}) are said to be **homeomorphic to each other** if there exists a map $f : X \rightarrow Y$ such that, (a) f is one-to-one and onto, and (b) both f and f^{-1} are continuous.³ Such a map f is called a **homeomorphism** from (X, \mathcal{T}) to (Y, \mathcal{S}) .

³ Motivated readers can try to find examples of continuous maps that are both one-to-one and onto whose inverse is discontinuous (hint: think about discrete and indiscrete topologies).

Fig. 1.5 $\forall x \in S^1$ (except for a), one can use the way shown in the figure to define its image $f(x)$ on \mathbb{R}



The continuity and differentiability of an ordinary function $y = f(x)$ can be represented by C^r , where r is a non-negative integer. C^0 stands for continuous, C^r indicates that the r th derivative exists and is continuous, and C^∞ denotes that derivatives of all orders exist and are continuous (called **smooth**). Although one can cleverly generalize the C^0 property to maps between topological spaces using the concept of open sets, this cannot be done for C^r with $r > 0$. In fact, the highest requirement for maps between topological spaces has already been reflected in the definition of homeomorphism. A homeomorphism $f : X \rightarrow Y$ not only sets up a one-to-one correspondence between the points in X and Y , but also sets up a one-to-one correspondence between the open sets of X and Y ; hence, all of the properties that depend on the topology can be “carried” into Y by f . Therefore, from a purely topological point of view, two topological spaces that are homeomorphic to each other “cannot be more alike”, and can be considered to be equivalent.

Example 6 Any open interval $(a, b) \subset \mathbb{R}$ is homeomorphic to \mathbb{R} (the proof is left to the reader as Exercise 1.6).

Example 7 A circle $S^1 \subset \mathbb{R}^2$ together with the induced topology (induced by \mathcal{T}_u of \mathbb{R}^2) can be treated as a topological space. Is it homeomorphic to \mathbb{R} ? At first glance, one might think it is possible to define a homeomorphism from S^1 to \mathbb{R} in terms of Fig. 1.5. However, f is not a map from S^1 to \mathbb{R} in that $a \in S^1$ has no image. It is not difficult to show that $f : (S^1 - \{a\}) \rightarrow \mathbb{R}$ is a homeomorphism; thus, a circle with a point removed is homeomorphic to \mathbb{R} . However, S^1 is not homeomorphic to \mathbb{R} ; we will give a concise proof after Theorem 1.3.8 in Sect. 1.3 (optional reading) in which the concept of “compactness” that will be discussed in Sect. 1.3 is used. The key points are: ① S^1 is compact but \mathbb{R} is not; ② The image of a compact subset under a continuous map is still compact. ① and ② imply that S^1 cannot be homeomorphic to \mathbb{R} .

Example 8 Consider a circle and an ellipse on a Euclidean plane. From the viewpoint of Euclidean geometry they are certainly different: the Euclidean geometry has a concept of distance which circles and ellipses are defined in respect to. However, from the aspect of pure topology, $(\mathbb{R}^2, \mathcal{T}_u)$ is a topological space and a circle S^1 as well as an ellipse E are two subsets of \mathbb{R}^2 : $S^1, E \subset \mathbb{R}^2$. One can make S^1 and E topological spaces (S^1, \mathcal{S}_{S^1}) and (E, \mathcal{S}_E) , where \mathcal{S}_{S^1} and \mathcal{S}_E are topologies induced by \mathcal{T}_u . It can be proved (and it is intuitively easy to believe) that there exists a homeomorphism $f : (S^1, \mathcal{S}_{S^1}) \rightarrow (E, \mathcal{S}_E)$; thus, from the perspective of

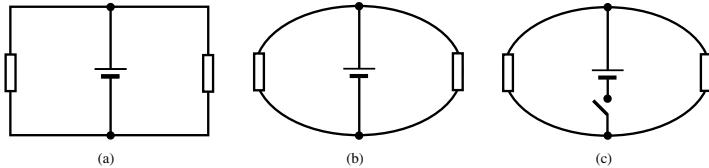


Fig. 1.6 From the perspective of circuits or topology, **a** and **b** are identical, while **b** and **c** are different. From the viewpoint of geometry, all three are different

pure topology they are exactly the same. Conversely, if we cut a gap on S^1 , the result will be homeomorphic to \mathbb{R} , and thus has a different topology from that of S^1 and E . If we imagine \mathbb{R}^2 as a rubber sheet and manipulate it by deformation, then the shape of a curve on the sheet will change with it. However, as long as there is no cutting or gluing, the curves before and after the deformation are homeomorphic to each other. Hence, topology is also colloquially called “rubber sheet geometry”. The major difference between topology geometry and Euclidean geometry is that the former does not have a concept of distance. At first glance, it may seem that a geometry without a concept of distance would not be useful, but this is not the case. A simple example is the electric circuit problem. Although there is a big difference between Fig. 1.6a, b from the Euclidean point of view, they are identical as circuits. Conversely, if we cut one of the branches in (b) [turning it to (c)], it will be very different in the view of circuits. This is the same as the viewpoint of topology. In fact, topology is very useful in the study of complex circuits (networks), which forms an applied branch called “network topology”.

Definition 5 $N \subset X$ is called a **neighborhood** of $x \in X$ if $\exists O \in \mathcal{T}$ such that $x \in O \subset N$. A neighborhood that is an open set is called an **open neighborhood**.

Remark 2 Suppose $X = \mathbb{R}$, and $N = [a, b]$, then N is a neighborhood of x according to Definition 5 if and only if $a < x < b$. Please pay particular attention to the “sideswipe” case: if $x = a$, then N is not a neighborhood of x , since there is no open set O in \mathbb{R} such that $x \in O \subset N$. Intuitively, to make $[a, b]$ a neighborhood of x , x should have “neighbors” on both sides of it. Since none of the “neighbors” on the left side belongs to $[a, b]$, $[a, b]$ cannot be a neighborhood of $x = a$. Thus, Definition 5 reflects this intuitive requirement to some extent. Please also note the following subtle example: In the topological space $[0, \infty] \subset \mathbb{R}$, an interval $[0, 1]$ is an open neighborhood of 0, while $[0, 1]$ is a neighborhood of 0.

Definition 5' (*Neighborhoods of a Subset*) $N \subset X$ is called a **neighborhood** of $A \subset X$ if $\exists O \in \mathcal{T}$ such that $A \subset O \subset N$.

Theorem 1.2.1 $A \subset X$ is open if and only if A is a neighborhood of $x \forall x \in A$.

Proof (A) Suppose A is open, then $\forall x \in A, \exists O \in \mathcal{T}$ such that $x \in O \subset A$. Hence, A is a neighborhood of x according to Definition 5.

(B) Suppose A is a neighborhood of $x \forall x \in A$, and let $O = \bigcup_{x \in A} O_x$ ($O_x \in \mathcal{T}$ satisfies $x \in O_x \subset A$ in Definition 5), then $O = A$ (the reader should complete the proof of this). Also, from Definition 1 (c) we know that $O \in \mathcal{T}$. Hence, $A \in \mathcal{T}$, i.e., A is an open set. \square

Definition 6 $C \subset X$ is called a **closed set** if $-C \in \mathcal{T}$.

Theorem 1.2.2 *Closed sets have the following properties:*

- (a) *The intersection of any number of closed sets is a closed set;*
- (b) *The union of a finite number of closed sets is a closed set;*
- (c) X and \emptyset are closed sets.

Proof They can be easily proved using Definitions 1, 6 and De Morgan's Law. \square

Thus, any topological space (X, \mathcal{T}) has two subsets that are both open and closed, namely X and \emptyset .

Definition 7 A topological space (X, \mathcal{T}) is said to be **connected** if it does not contain a subset that is both open and closed other than X and \emptyset .

Example 9 Suppose A and B are open intervals of \mathbb{R} , and $A \cap B = \emptyset$ (draw a picture of this). If we use \mathcal{T} to represent the topology induced on the subset $X \equiv A \cup B$ by the usual topology of \mathbb{R} , then, in addition to X and \emptyset , the topological space (X, \mathcal{T}) also has subsets A and B that are both open and closed (A and B are open under the induced topology, and they are also closed since they are complements of one another.). Thus, (X, \mathcal{T}) is not connected, which coincides with the fact that the picture of A and B you drew is intuitively not connected.⁴

Suppose (X, \mathcal{T}) is a topological space, and let $A \subset X$. The closure, interior and boundary of A are defined as follows:

Definition 8 The **closure** of A , denoted by \bar{A} , is the intersection of all of the closed sets that contain A , i.e.,

$$\bar{A} := \bigcap_{\alpha} C_{\alpha}, \quad A \subset C_{\alpha}, \quad \text{and } C_{\alpha} \text{ is closed.} \quad (1.2.3)$$

Definition 9 The **interior** of A , denoted by $i(A)$, is the union of all open sets that are contained in A , i.e.,

$$i(A) := \bigcup_{\alpha} O_{\alpha}, \quad O_{\alpha} \subset A, \quad O_{\alpha} \in \mathcal{T}. \quad (1.2.4)$$

Definition 10 The **boundary** of A is defined as $\dot{A} := \bar{A} - i(A)$. $x \in \dot{A}$ is called a **boundary point**. \dot{A} is also denoted by ∂A .

⁴ What is more consistent with our intuition is the concept called “arcwise connected”. There are subtle differences between this and the concept of connected (see the first footnote of Sect. 5.2).

Theorem 1.2.3 \bar{A} , $i(A)$ and \dot{A} have the following properties:

- (a) ① \bar{A} is a closed set, ② $A \subset \bar{A}$, ③ $A = \bar{A}$ if and only if A is a closed set;
- (b) ① $i(A)$ is an open set, ② $i(A) \subset A$, ③ $i(A) = A$ if and only if $A \in \mathcal{T}$;
- (c) \dot{A} is a closed set.

Proof (a), (b) are easy to prove. (c) can be proved as follows: $X - \dot{A} = X - [\bar{A} - i(A)] = (X - \bar{A}) \cup i(A)$, where we used the conclusion of Exercise 1.2 in the last step. Since \bar{A} is closed, $X - \bar{A}$ is open. In addition, $i(A)$ is open, so hence have $X - \dot{A}$ is open. Therefore, \dot{A} is closed. \square

The definition below will be used in the whole of Sect. 1.3 and the beginning of Chap. 2:

Definition 11 A set $\{O_\alpha\}$ of open sets of X is called an **open cover** of $A \subset X$ if $A \subset \bigcup_\alpha O_\alpha$. We can also say that $\{O_\alpha\}$ covers A .

1.3 Compactness [Optional Reading]

Definition 1 Suppose $\{O_\alpha\}$ is an open cover of $A \subset X$. If $\{O_{\alpha_1}, \dots, O_{\alpha_n}\}$, a subset of $\{O_\alpha\}$ with finitely many elements, also covers A , then we say $\{O_\alpha\}$ has a **finite subcover**.

Definition 2 $A \subset X$ is said to be **compact** if any of its open covers has a finite subcover.

Example 1 Suppose $x \in X$, then the one-point subset $A \equiv \{x\}$ must be compact.

Proof Suppose $\{O_\alpha\}$ is an arbitrary open cover of A , then there exists at least one element in $\{O_\alpha\}$ (denoted by $\{O_{\alpha_1}\}$) that satisfies $x \in \{O_{\alpha_1}\}$. Hence, $\{O_{\alpha_1}\}$ (as a subset of $\{O_\alpha\}$) is an open cover of $A \equiv \{x\}$, and thus $\{O_\alpha\}$ has a finite subcover. \square

Example 2 $A \equiv (0, 1] \subset \mathbb{R}$ is not compact.

Proof Let \mathbb{N} represent the set of natural numbers, then $\{(1/n, 2) \mid n \in \mathbb{N}\}$ is an open cover of A that does not have finite subcover. \square

Similarly, any open interval or half-open interval in \mathbb{R} is noncompact.

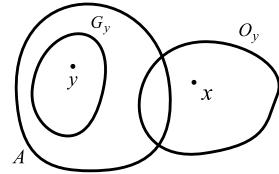
Example 3 \mathbb{R} is not compact (the proof is left as Exercise 1.16).

Theorem 1.3.1 Any closed interval of \mathbb{R} is compact.

Proof Omitted. \square

Remark 1 Do not think that closed sets are necessarily compact (even \mathbb{R} has noncompact closed subsets; the reader should try to find an example). Compactness and closedness are closely related, but not equivalent. Their relationship is shown in the following two theorems.

Fig. 1.7 Figure for the proof of Theorem 1.3.2



To give Theorem 1.3.2, we first introduce the following definition.

Definition 3 A topological space (X, \mathcal{T}) is called a **T₂ space** or **Hausdorff space** if $\forall x, y \in X, x \neq y, \exists O_1, O_2 \in \mathcal{T}$ such that $x \in O_1, y \in O_2$ and $O_1 \cap O_2 = \emptyset$.

Remark 2 Almost all of the common topological spaces (such as \mathbb{R}^n) are T₂ spaces. The indiscrete topological space is an example of a non-T₂ space. Hawking and Ellis (1973) (pp. 13–14) provided an example that is “closer to practical use”.

Theorem 1.3.2 If (X, \mathcal{T}) is a T₂ space, $A \subset X$ is compact, then A is a closed set.

Proof The theorem obviously holds when $A = \emptyset$, and thus we suppose $A \neq \emptyset$ below. All we have to prove is that $X - A \in \mathcal{T}$; to prove this we only have to show that $\forall x \in X - A, \exists O \in \mathcal{T}$ such that $x \in O \subset X - A$ (see Theorem 1.2.1). Since X is a T₂ space, when x is given, $\forall y \in A, \exists O_y, G_y \in \mathcal{T}$ such that $x \in O_y, y \in G_y$ and $O_y \cap G_y = \emptyset$ (see Fig. 1.7). Varying y over A yields two sets of subsets $\{G_y \mid y \in A\}$ and $\{O_y \mid y \in A\}$. It is easy to see that $\{G_y \mid y \in A\}$ is an open cover of A . The compactness of A assures that it must contain a finite subcover $\{G_{y_1}, \dots, G_{y_n}\}$. Let $O \equiv O_{y_1} \cap \dots \cap O_{y_n}$, then we have: ① $O \in \mathcal{T}$; ② $x \in O$; ③ $O \cap A = \emptyset$ (the proof is left as an exercise), i.e., $O \subset X - A$. Thus, from Theorem 1.2.1 we know that $X - A \in \mathcal{T}$, and hence A is closed. \square

Theorem 1.3.3 If (X, \mathcal{T}) is compact and $A \subset X$ is closed, then A is compact.

Proof Since A is closed, $X - A$ is open. Suppose $\{O_\alpha\}$ is an arbitrary open cover of A , then $\{O_\alpha, X - A\}$ is an open cover of X (here we used the fact that A is a closed set). X is compact indicates that there exists a finite subcover $\{O_1, \dots, O_n; X - A\}$ for $\{O_\alpha, X - A\}$. Therefore, $\{O_1, \dots, O_n\}$ covers A , and hence $\{O_\alpha\}$ has a finite subcover. \square

Definition 4 $A \subset \mathbb{R}^n$ is said to be **bounded** if \exists an open ball $B \subset \mathbb{R}^n$ such that $A \subset B$.

Theorem 1.3.4 $A \subset \mathbb{R}$ is compact if and only if A is a bounded closed set.

Proof (A) Suppose A is compact. (a) Since \mathbb{R} is a T₂ space, from Theorem 1.3.2 we know that A is a closed set. (b) $\{(-n, n) \mid n \in \mathbb{N}\}$ is an open cover of A , the compactness of A assures that for this open cover there exists a finite subcover $\{(-1, 1), (-2, 2), \dots, (-m, m)\}$, i.e., $A \subset (-1, 1) \cup (-2, 2) \cup \dots \cup (-m, m) = (-m, m)$, and thus A is bounded.

(B) Suppose A is a bounded closed set. The boundedness assures that $\exists M \in \mathbb{R}$ such that $A \subset [-M, M]$. From Theorem 1.3.1 we know that $[-M, M]$, being a subset of $(\mathbb{R}, \mathcal{T}_u)$, is compact. Let $C \equiv [-M, M]$ and use \mathcal{S} to represent the topology induced on C by \mathcal{T}_u , then it can be proved that (C, \mathcal{S}) is also compact (exercise). Regarding (C, \mathcal{S})

as (X, \mathcal{T}) in Theorem 1.3.3 and noticing that $A \subset C$ is closed, we conclude that A is compact.⁵ \square

Theorem 1.3.5 Suppose $A \subset X$ is compact, and $f : X \rightarrow Y$ is continuous, then $f[A] \subset Y$ is compact.

Proof Suppose $\{O_\alpha\}$ is an arbitrary open cover of $f[A]$. The continuity of f assures that $f^{-1}[O_\alpha]$ is open, and thus $\{f^{-1}[O_\alpha]\}$ is an open cover of A . Since A is compact, there exists a finite subcover $\{f^{-1}[O_1], \dots, f^{-1}[O_n]\}$; thus, $\{O_1, \dots, O_n\}$ is an open subcover of $\{O_\alpha\}$. Therefore, $f[A] \subset Y$ is compact. \square

From Theorem 1.3.5 we can obtain a corollary: homeomorphisms preserve the compactness of subsets.

Definition 5 A property that is invariant under homeomorphisms is called a **topological property** or **topological invariance**.

Example 4 Compactness, connectedness, and the property of T_2 are all topological properties. Boundedness is not an topological property; for example, although an interval (a, b) is homeomorphic to \mathbb{R} , the former is bounded while the latter is unbounded. From this it can also be seen that length is not a topological property either.

There is a well-known theorem in mathematical analysis: Any continuous function on a closed interval must attain its maximum and minimum value on this interval. The following theorem is a generalization of this theorem.

Theorem 1.3.6 Suppose X is compact and $f : X \rightarrow \mathbb{R}$ is continuous, then $f[X] \subset \mathbb{R}$ is bounded and attains a maximum and minimum value.

Proof This is a corollary of Theorems 1.3.4 and 1.3.5. \square

Theorem 1.3.7 Suppose $(X, \mathcal{T}_1), (Y, \mathcal{T}_2)$ are compact, then $(X \times Y, \mathcal{T})$ is compact (\mathcal{T} is the product topology of \mathcal{T}_1 and \mathcal{T}_2).

Proof Omitted. \square

Theorem 1.3.8 $A \subset \mathbb{R}^n$ is compact if and only if it is a bounded closed set.

Proof This is a corollary of Theorem 1.3.7 and previous theorems (\mathbb{R}^n is the Cartesian product of $n \mathbb{R}$). \square

Simple Application Example. Consider $(\mathbb{R}^2, \mathcal{T}_0)$. Suppose S^1 is an arbitrary circle in \mathbb{R}^2 , then it is easy to see that it is a bounded closed set; thus, from Theorem 1.3.8 we know that it is compact. From Theorem 1.3.5 we know that continuous maps preserve compactness; however, neither \mathbb{R} nor any of its open intervals is compact. Therefore, S^1 cannot be homeomorphic to \mathbb{R} or any of its open intervals. Similarly, from Theorems 1.3.1 and 1.3.5 we know that any closed interval cannot be homeomorphic to \mathbb{R} or any of its open intervals.

⁵ Strictly speaking, to get the conclusion that A is compact from Theorem 1.3.3 we should generalize this theorem slightly as follows (its proof is similar to the original theorem): suppose C is a compact subset of a topological space (X, \mathcal{T}) , $A \subset C$ and A is a closed subset of (X, \mathcal{T}) , then A must be compact.

Definition 6 A map $S : \mathbb{N} \rightarrow X$ is called a **sequence** of X .

Remark 3 Usually a sequence is denoted by $\{x_n\}$, where $x_n \equiv S(n) \in X$, $n \in \mathbb{N}$. $\{x_n\}$ is actually just a series of ordered points in X .

Definition 7 $x \in X$ is called the **limit** of a sequence $\{x_n\}$ if for any open neighborhood O of x there exists $N \in \mathbb{N}$ such that $x_n \in O \forall n > N$. If x is the limit of $\{x_n\}$, then we say $\{x_n\}$ **converges** to x .

Definition 8 $x \in X$ is called an **accumulation point** of a sequence $\{x_n\}$ if any open neighborhood of x contains infinitely many points of $\{x_n\}$.

Remark 4 x is the limit of $\{x_n\} \Rightarrow x$ is an accumulation point of $\{x_n\}$, but not vice versa.

One of the conditions in the following theorem involves the concept “second countable”. A set with a finite number of elements is called a **finite set**; otherwise; it is called an **infinite set**. For a finite set one can always number its elements and count them one by one, so a finite set must be a countable set. However, an infinite set is not necessarily uncountable; for example, \mathbb{N} is a countable infinite set. Finite sets are simpler than infinite sets, and countable infinite sets are simpler than uncountable infinite sets. A topological space (X, \mathcal{T}) is said to be **second countable** if there exists a countable subset $\{O_1, \dots, O_K\} \subset \mathcal{T}$ or $\{O_1, \dots\} \subset \mathcal{T}$ for \mathcal{T} such that any $O \in \mathcal{T}$ can be expressed as a union of elements from $\{O_1, \dots, O_K\}$ or $\{O_1, \dots\}$. For example, $(\mathbb{R}^n, \mathcal{T}_u)$ is second countable since \mathcal{T}_u has a countable subset such that any $O \in \mathcal{T}_u$ can be expressed as a union of elements from this subset. (This countable subset above is a subset of \mathcal{T}_u such that each element O_i of it is an open ball, the natural coordinates of its center are all rational numbers, so is its radius.)

Theorem 1.3.9 If $A \subset X$ is compact, then any sequence in A has an accumulation point within A . Conversely, if X is second countable and any sequence in $A \subset X$ has an accumulation point within A , then A is compact.

Proof Omitted. □

Exercises

- ~1.1. Show that $A - B = A \cap (X - B)$, $\forall A, B \in X$.
 - ~1.2. Show that $X - (B - A) = (X - B) \cup A$, $\forall A, B \in X$.
 - ~1.3. Fill in the following table with “True” or “False”:
- | $f : \mathbb{R} \rightarrow \mathbb{R}$ | is a one-to-one map | is an onto map |
|---|---------------------|----------------|
| $f(x) = x^3$ | | |
| $f(x) = e^x$ | | |
| $f(x) = \cos x$ | | |
| $f(x) = 5, \forall x \in \mathbb{R}$ | | |
- ~1.4. Determine whether each of the following statements is true or false, and give a brief explanation:

- (a) the tangent function is a map from $\mathbb{R} \rightarrow \mathbb{R}$;
 - (b) a logarithmic function is a map from $\mathbb{R} \rightarrow \mathbb{R}$;
 - (c) $(a, b] \subset \mathbb{R}$ is an open set measured by \mathcal{T}_u ;
 - (d) $[a, b] \subset \mathbb{R}$ is a closed set measured by \mathcal{T}_u .
- ~1.5. Give a counterexample to show that the statement “the intersection of an infinite number of open subsets of $(\mathbb{R}, \mathcal{T}_u)$ is open” is not true.
- ~1.6. Show that the induced topology defined in Example 5 of Sect. 1.2 satisfies the three conditions in Definition 1 of that section.
- 1.7. Use an example to show that $(\mathbb{R}^3, \mathcal{T}_u)$ has a subset that is neither open nor closed.
- ~1.8. Is the constant map $f : (X, \mathcal{T}) \rightarrow (Y, \mathcal{S})$ continuous? Why?
- ~1.9. Suppose \mathcal{T} is a discrete topology on a set X , and \mathcal{S} is an indiscrete topology on a set Y .
 - (a) Find all the continuous maps from (X, \mathcal{T}) to (Y, \mathcal{S}) . (b) Find all the continuous maps from (Y, \mathcal{S}) to (X, \mathcal{T}) .
- ~1.10. Show that Definitions 3a and 3b are equivalent.
- 1.11. Show that any open interval $(a, b) \in \mathbb{R}$ is homeomorphic to \mathbb{R} .
- 1.12. Suppose X_1 and X_2 are subsets of \mathbb{R} , $X_1 \equiv (1, 2) \cup (2, 3)$, $X_2 \equiv (1, 2) \cup [2, 3]$. The topologies induced on X_1 and X_2 by the usual topology of \mathbb{R} , respectively, are denoted by \mathcal{T}_1 and \mathcal{T}_2 . Are the topological spaces (X_1, \mathcal{T}_1) and (X_2, \mathcal{T}_2) connected?
- 1.13. Is the topological space formed by a set X with the discrete topology \mathcal{T} connected?
- ~1.13. Suppose $A \subset B$. Show that (a) $\bar{A} \subset \bar{B}$. Hint: $A \subset B$ implies that \bar{B} is a closed set that contains A . (b) $i(A) \in i(B)$.
- ~1.14. Show that $x \in \bar{A} \Leftrightarrow$ the intersection of A and any neighborhood of x is non-empty. A hint to \Rightarrow : suppose O in \mathcal{T} and $O \cap A = \emptyset$. First show that $A \subset X - O$, then (use the definition of closure) show that $\bar{A} \in X - O$.
- 1.14. Show that \mathbb{R} is not compact.

Reference

Hawking, S. W. and Ellis, G. F. R. (1973), *The Large Scale Structure of Space-Time*, Cambridge University Press, Cambridge.

Chapter 2

Manifolds and Tensor Fields



2.1 Differentiable Manifolds

Physics cannot be done without a background space. For example, classical mechanics and electrodynamics study the time evolution of matter and electromagnetic fields in \mathbb{R}^3 , statistical physics and Hamiltonian theory often use phase spaces, special relativity has \mathbb{R}^4 as its spacetime background, etc. Colloquially, these spaces are all “continuous” rather than consisting of discrete points. The spacetime of general relativity is also a “continuous 4-dimensional space”, which locally looks like \mathbb{R}^4 , yet is not necessarily \mathbb{R}^4 . However, the meaning of the word “continuous” is not yet clear. “Differentiable manifold” (or “manifold” for short) is the accurate term used for all kinds of “continuous spaces” with differential structures. \mathbb{R}^n is the simplest n -dimensional manifold. Roughly speaking, differential manifolds are topological spaces with differential structures, which look locally like \mathbb{R}^n , but globally may be different from \mathbb{R}^n . The precise definition is as follows:

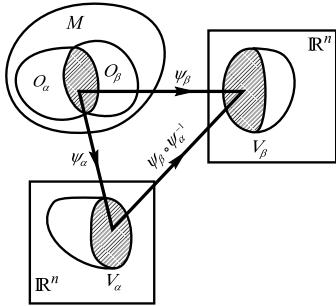
Definition 1 A topological space M is called an **n -dimensional differentiable manifold**, or n -dimensional manifold for short, if M has an open cover $\{O_\alpha\}$, i.e., $M = \bigcup_\alpha O_\alpha$ (see Definition 11 of Sect. 1.2), satisfying

- (a) for each $O_\alpha \exists$ a homeomorphism $\psi_\alpha : O_\alpha \rightarrow V_\alpha$ (V_α is an open subset of \mathbb{R}^n measured by the usual topology);
- (b) If $O_\alpha \cap O_\beta \neq \emptyset$, then the composite map $\psi_\beta \circ \psi_\alpha^{-1}$ (see Fig. 2.1) is C^∞ (smooth).¹

Remark 1 ① $\psi_\beta \circ \psi_\alpha^{-1}$ is a map from $\psi_\alpha[O_\alpha \cap O_\beta] \subset \mathbb{R}^n$ to $\psi_\beta[O_\alpha \cap O_\beta] \subset \mathbb{R}^n$. Since each point of \mathbb{R}^n has n natural coordinates, $\psi_\beta \circ \psi_\alpha^{-1}$ provides n functions of n variables (see Remark 4 of Sect. 1.1). “ $\psi_\beta \circ \psi_\alpha^{-1}$ is C^∞ ” means that all these functions of n variables are C^∞ (the smoothness of n -variable functions has already

¹ Definition 1 is the general definition of a smooth manifold. In this text, and usually in physics, manifolds also satisfy the following additional conditions: as a topological space, M is Hausdorff and second countable (for both see Sect. 1.3). From now on, our manifolds will satisfy these conditions.

Fig. 2.1 Figure for the definition of a manifold. $\psi_\beta \circ \psi_\alpha^{-1}$ is the composite map of ψ_α^{-1} and ψ_β



been defined in calculus).² ② Suppose $p \in O_\alpha$, then $\psi_\alpha(p) \in \mathbb{R}^n$, and thus the point $\psi_\alpha(p)$ has n natural coordinates. It is natural to call these n numbers the **coordinates** of p acquired through the map ψ_α . Being a topological space, M in general does not have coordinates; being a manifold, however, its elements (points) in O_α can acquire coordinates from the map ψ_α . If $O_\alpha \cap O_\beta \neq \emptyset$, then the points in $O_\alpha \cap O_\beta$ can acquire coordinates from either ψ_α or ψ_β , and these two sets of coordinates are different in general. We say that (O_α, ψ_α) forms a (local) **coordinate system** whose **coordinate patch** is O_α ; (O_β, ψ_β) forms another coordinate system whose coordinate patch is O_β . Thus, a point in $O_\alpha \cap O_\beta$ has at least two sets of coordinates, denoted by $\{x^\mu\}$ and $\{x^\nu\}$ ($\mu, \nu = 1, \dots, n$), respectively. The map $\psi_\beta \circ \psi_\alpha^{-1}$ naturally provides a relation connecting these two sets of coordinates, which is represented by n functions of n variables as follow:

$$x'^1 = \phi^1(x^1, \dots, x^n), \quad \dots, \quad x'^n = \phi^n(x^1, \dots, x^n).$$

This relation is called a **coordinate transformation**. Condition (b) in Definition 1 assures that the function relations $x'^\mu = \phi^\mu(x^1, \dots, x^n)$ in a coordinate transformation are all C^∞ . For convenience's sake, $\{x^\nu\}$ is also usually called a coordinate system, although we cannot see the domain of the coordinate patch explicitly from $\{x^\nu\}$. Physicists also often denote $x'^\mu = \phi^\mu(x^1, \dots, x^n)$ as $x'^\mu = x'^\mu(x^1, \dots, x^n)$.

Definition 2 In mathematics, a coordinate system (O_α, ψ_α) is also called a **chart**; the collection of all charts $\{(O_\alpha, \psi_\alpha)\}$ satisfying conditions (a) and (b) in Definition 1 is called an **atlas**. Condition (b) is also called the **compatibility condition**. Therefore, any two charts in an atlas are compatible.

Example 1 Suppose $M = (\mathbb{R}^2, \mathcal{T}_0)$. Choose $O_1 = \mathbb{R}^2$, ψ_1 = identity map (i.e., each image coincides with its inverse image), then $\{(O_1, \psi_1)\}$ is an atlas that contains only one chart. Thus, \mathbb{R}^2 is a 2-dimensional manifold that can be covered by a single coordinate patch, called a **trivial manifold**. According to this atlas, the coordinates of each point in \mathbb{R}^2 are the natural coordinates it is endowed with as an element of

² The manifold in Definition 1 is short for **smooth manifold**. Change C^∞ in condition (b) to C^r (r is a natural number), then it becomes the definition of a C^r manifold.

\mathbb{R}^2 . Of course points in \mathbb{R}^2 can also be described by other coordinates (such as polar coordinates). This is actually nothing but choosing another chart (O_2, ψ_2) that is compatible with (O_1, ψ_1) , where ψ_2 maps $p \in O_2$ to $\psi_2(p) \in \mathbb{R}^2$, and referring to the natural coordinates of $\psi_2(p)$ as the new coordinates of p . However, it should be noted that the coordinate patch of \mathbb{R}^2 does not necessarily include all points of \mathbb{R}^2 (e.g., polar coordinates).

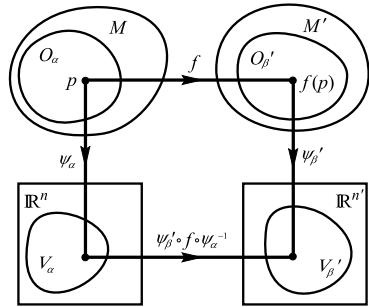
Similarly, we can see that \mathbb{R}^n is an n -dimensional trivial manifold.

Example 2 Suppose $M = (S^1, \mathcal{S})$, where $S^1 := \{x \in \mathbb{R}^2 \mid |x - o| = 1\}$ is a unit circle with center o and \mathcal{S} is the topology on S^1 induced by \mathcal{T}_u of \mathbb{R}^2 . Since S^1 is not homeomorphic to \mathbb{R} (see Example 7 in Sect. 1.2), any atlas that defines (S^1, \mathcal{S}) as a manifold cannot contain only one chart. Suppose x^1 and x^2 are the natural coordinates of \mathbb{R}^2 , and $O_1^+, O_1^-, O_2^+, O_2^-$ are “open semicircles” defined as follows: $O_i^+ = \{(x^1, x^2) \in S^1 \mid x^i > 0\}$, $O_i^- = \{(x^1, x^2) \in S^1 \mid x^i < 0\}$, $i = 1, 2$, then $\{O_i^\pm\}$ can cover S^1 . Define the homeomorphism ψ_i^\pm from O_i^\pm to the open interval $(-1, 1)$ as the following “projection map”: $\psi_1^\pm((x^1, x^2)) = x^2$, $\psi_2^\pm((x^1, x^2)) = x^1$, then it is easy to prove that the overlap regions of the open semicircles satisfy the compatibility condition (Exercise 2.1), and thus S^1 is a 1-dimensional manifold. In fact, one can cover S^1 with an atlas containing only two charts; motivated readers may try to verify that.

Example 3 Suppose $M = (S^2, \mathcal{T})$, where $S^2 = \{x \in \mathbb{R}^3 \mid |x - o| = 1\}$ is a unit sphere with center o and \mathcal{S} is the topology on S^2 induced by \mathcal{T}_u of \mathbb{R}^3 . Following the method in the last example, one can cover S^2 with six open hemispheres and define the homeomorphism from each hemisphere to the corresponding open disk on \mathbb{R}^2 [Wald (1984)]. By showing that the overlap regions satisfy the compatibility condition one can show that S^2 is a 2-dimensional manifold. It can also be proved that S^2 can be covered with an atlas that contains only two charts. The surface of the Earth can be regarded as S^2 , while it looks locally like \mathbb{R}^2 . You cannot tell that human beings are living on a sphere from just a city map of Beijing (\mathbb{R}^2). On the contrary, a globe tells you clearly that the surface of the Earth is globally not \mathbb{R}^2 .

Suppose the atlas $\{(O_\alpha, \psi_\alpha)\}$ defines a topological space M as a manifold, then any two charts in this atlas are naturally compatible. However, we can also use another atlas $\{(O'_\beta, \psi'_\beta)\}$ to define the same M as a manifold. Now there are two possibilities: ① These two atlases are not compatible with each other, i.e., there exists O_α and O'_β such that $O_\alpha \cap O'_\beta \neq \emptyset$, and ψ_α and ψ'_β do not satisfy condition (b) of Definition 1 on $O_\alpha \cap O'_\beta$. We then say that these two atlases define M as two different differentiable manifolds, and these two atlas represent two different **differential structures** (The concept of differential structures should be understood gradually, rather than in one step.); ② These two atlases are compatible, then we say they define M as the same differentiable manifold (with only one differential structure). For convenience, we may treat $\{(O_\alpha, \psi_\alpha); (O'_\beta, \psi'_\beta)\}$ as one atlas. Furthermore, we may even put all of the charts compatible with (O_α, ψ_α) together and create one large atlas. Later on, when we talk about a manifold, we always assume that the largest possible atlas

Fig. 2.2 The map $\psi'_\beta \circ f \circ \psi_\alpha^{-1}$ corresponds to n' functions of n variables, whose C^r -differentiability defines the C^r -differentiability of $f : M \rightarrow M'$



has been chosen as the differential structure, so that we can perform any coordinate transformation.

A significant difference between differentiable manifolds and topological spaces is that the former has differential structures additional to topological structures. Therefore, for a map between two manifolds we can not only talk about whether it is continuous, but also whether it is differentiable, or even if it is C^∞ . Suppose M and M' are two manifolds whose dimensions are n and n' , atlases are $\{(O_\alpha, \psi_\alpha)\}$ and $\{(O'_\beta, \psi'_\beta)\}$, respectively, and $f : M \rightarrow M'$ is a continuous map (see Fig. 2.2). $\forall p \in M$, choose any coordinate system (O_α, ψ_α) such that $p \in O_\alpha$, and coordinate system (O'_β, ψ'_β) such that $f(p) \in O'_\beta$, then $\psi'_\beta \circ f \circ \psi_\alpha^{-1}$ is a map from $V_\alpha \equiv \psi_\alpha[O_\alpha]$ (or an open set of V_α) to $\mathbb{R}^{n'}$. Thus, this map corresponds to n' functions of n variables, and their C^r -differentiability can be used to define the C^r -differentiability for $f : M \rightarrow M'$.

Definition 3 $f : M \rightarrow M'$ is called a **C^r map** if $\forall p \in M$, the n' functions of n variables corresponding to the map $\psi'_\beta \circ f \circ \psi_\alpha^{-1}$ are of class C^r .

Remark 2 Since charts in the same atlas are compatible, the definition above is independent of the choice of (O_α, ψ_α) and (O'_β, ψ'_β) .

Definition 4 Differential manifolds M and M' are said to be **diffeomorphic to each other** if $\exists f : M \rightarrow M'$ satisfying (a) f is one-to-one and onto; (b) f and f^{-1} are C^∞ . Such an f is called a **diffeomorphism** from M to M' .

Remark 3 ① Being a diffeomorphism is the highest requirement one can impose on a map of manifolds (if there are additional structures imposed on these manifolds then it is another matter); manifolds that are diffeomorphic to each other can be considered to be equivalent. ② A necessary condition for two manifolds to be diffeomorphic to each other is that they have the same dimension. ③ In Definition 1, $\psi : O_\alpha \rightarrow V_\alpha$ was required to be a homeomorphism instead of a diffeomorphism since a diffeomorphism is a relationship between manifolds, and we did not have the concept of manifold yet. But now that Definition 4 has been introduced, one may naturally ask: if we treat O_α and V_α in Definition 1 as manifolds, is ψ_α a diffeomorphism? The answer is affirmative, and motivated readers should try to verify this. From this, one can further their understanding of the statement “a manifold M looks locally like \mathbb{R}^n ”.

A simple, but important, example of a map $f : M \rightarrow M'$ is the case $M' = \mathbb{R}$. In this case each point of M corresponds to a real number, and hence we have the following definition:

Definition 5 $f : M \rightarrow \mathbb{R}$ is called a **function on M** or a **scalar field on M** . If f is C^∞ , then it is called a **smooth function on M** . The collection of all smooth functions on M is denoted by \mathcal{F}_M , abbreviated with \mathcal{F} when there is no confusion. From now on, functions will refer to smooth functions unless stated otherwise.

Example 4 The electric potential of a point charge at a point q in \mathbb{R}^3 is a smooth function on the manifold $M \equiv \mathbb{R}^3 - \{q\}$.

Example 5 The μ th coordinate x^μ of a coordinate system (O, ψ) is a smooth function defined on O ; interested readers may try to show that it satisfies the definition of a smooth function.

Since n coordinates determine a unique point p in O and from $f : M \rightarrow \mathbb{R}$ we get a unique real number $f(p)$, when a function $f : M \rightarrow \mathbb{R}$ is combined with a coordinate system (O, ψ) , we get a function of n variables $F(x^1, \dots, x^n)$. However, when f is combined with another coordinate system (O, ψ') , we have another function of n variables $F'(x'^1, \dots, x'^n)$. The function relations F and F' are different since $F = f \circ \psi^{-1}$ while $F' = f \circ \psi'^{-1}$. Thus, the multivariate function (function relation) corresponding to $f : M \rightarrow \mathbb{R}$ is coordinate dependent. One should pay attention to distinguishing a function (scalar field) f from the multivariate function which comes from combining f with a coordinate system.

Suppose M, N are manifolds, then they must be topological spaces, and thus $M \times N$ is also a topological space. It is not difficult to go a step further and define $M \times N$ as a manifold using the manifold structures of M and N [see Wald (1984) p. 13]. Suppose M and N have dimensions m and n respectively, then the dimension of $M \times N$ is $m + n$, i.e., $\dim(M \times N) = \dim M + \dim N$.

2.2 Tangent Vectors and Tangent Vector Fields

2.2.1 Tangent Vectors

First we review the definition of a vector space (i.e., linear space) in linear algebra.

Definition 1 A **vector space** over the field of real numbers is a set V together with two maps, namely $V \times V \rightarrow V$ (called **addition**) and $\mathbb{R} \times V \rightarrow V$ (called **scalar multiplication**), satisfying the following conditions:

- (a) $v_1 + v_2 = v_2 + v_1, \quad \forall v_1, v_2 \in V;$
- (b) $(v_1 + v_2) + v_3 = v_1 + (v_2 + v_3), \quad \forall v_1, v_2, v_3 \in V;$
- (c) \exists a zero element $\underline{0} \in V$ such that $\underline{0} + v = v, \quad \forall v \in V;$
- (d) $\alpha_1(\alpha_2 v) = (\alpha_1 \alpha_2)v, \quad \forall v \in V, \quad \alpha_1, \alpha_2 \in \mathbb{R};$

- (e) $(\alpha_1 + \alpha_2)v = \alpha_1v + \alpha_2v, \quad \forall v \in V, \quad \alpha_1, \alpha_2 \in \mathbb{R};$
- (f) $\alpha(v_1 + v_2) = \alpha v_1 + \alpha v_2, \quad \forall v_1, v_2 \in V, \quad \alpha \in \mathbb{R};$
- (g) $1 \cdot v = v, \quad 0 \cdot v = \underline{0}, \quad \forall v \in V.$

Remark 1 From these 7 conditions we can deduce that: $\forall v \in V, \exists u \in V$ such that $v + u = \underline{0}$. u is conventionally denoted by $-v$.

We will also often denote the zero element of V as 0; that is, the symbol 0 stands for both $0 \in \mathbb{R}$ and $\underline{0} \in V$. The reader should be able to identify its meaning by the context or equation.

Algebraically speaking, any set that satisfies Definition 1 is called a vector space, and any element of it is called a vector. Suppose p is a point in 3-dimensional Euclidean space, and V_p is the collection of straight line segments (or arrows \vec{v}) that start at p with all possible directions and lengths. Define the addition of two arrows as adding them by the parallelogram law, and define the scalar multiplication $\alpha\vec{v}$ ($\forall \alpha \in \mathbb{R}, \vec{v} \in V_p$) as the manipulation that preserves the direction of the arrow (or turns into the opposite direction when $\alpha < 0$) while multiplying its length by $|\alpha|$, then V_p is a vector space according to Definition 1, and hence each arrow starting at p is a vector. We want to generalize this kind of concept of vectors to an arbitrary manifold M ; that is, we want to define infinitely many vectors at each point p of M , the collection of which forms a vector space at p . Since “straight line segment”, “direction” and “length” are not defined on a general manifold (yet), the way we defined vectors in terms of arrows cannot be carried over to general manifolds. To generalize, we should pick the most essential property of an arrow which is the easiest to be generalized. Suppose \vec{v} is an arrow at an arbitrary point p in \mathbb{R}^3 , then we can take the directional derivative of an arbitrary C^∞ function f on \mathbb{R}^3 along \vec{v} , and the value of this derivative function at p is a real number. Thus, \vec{v} is a map that turns f into a real number. Let $\mathcal{F}_{\mathbb{R}^3}$ represent the collection of all smooth functions on \mathbb{R}^3 , then $f \in \mathcal{F}_{\mathbb{R}^3}$, and hence \vec{v} is a map from $\mathcal{F}_{\mathbb{R}^3}$ to \mathbb{R} , i.e., $\vec{v} : \mathcal{F}_{\mathbb{R}^3} \rightarrow \mathbb{R}$. Since the manipulation of taking the directional derivative is linear and satisfies the Leibniz rule, we finally have found the essential property of an arrow \vec{v} that can be easily generalized: it is a linear map from $\mathcal{F}_{\mathbb{R}^3}$ to \mathbb{R} that satisfies the Leibniz rule. Generalizing this to an arbitrary point p of an arbitrary manifold M , we arrive at the following definition:

Definition 2 A map $v : \mathcal{F}_M \rightarrow \mathbb{R}$ is called a **vector at a point $p \in M$** if $\forall f, g \in \mathcal{F}_M, \alpha, \beta \in \mathbb{R}$ we have

- (a) (Linearity) $v(\alpha f + \beta g) = \alpha v(f) + \beta v(g);$
- (b) (Leibniz rule) $v(fg) = f|_p v(g) + g|_p v(f)$, where $f|_p$ stands for the value of the function f at p , which can also be denoted by $f(p)$.

Remark 2 Since f and g are functions on M , fg is also a function on M whose value at any point p of M is defined as the product of $f(p)$ and $g(p)$.

[Optional Reading 1.2.1]

Theorem 2.2.1 Suppose $f_1, f_2 \in \mathcal{F}_M$ are equal in a neighborhood N of $p \in M$, i.e., $f_1|_N = f_2|_N$, then for any vector v at p we have $v(f_1) = v(f_2)$.

Proof First we prove two lemmas.

Lemma 1 If $f \in \mathcal{F}_M$ is a zero function, i.e., $f|_p = 0 \forall p \in M$, then for any vector v at p we have $v(f) = 0$.

Proof of Lemma 1 $\forall g \in \mathcal{F}_M$ we have $f + g = g$. The linearity of the action of v yields

$$v(g) = v(f + g) = v(f) + v(g),$$

and hence $v(f) = 0$. \square

Lemma 2 If $f \in \mathcal{F}_M$ is zero in a neighborhood N of $p \in M$, i.e., $f|_N = 0$, then for any vector v at p we have $v(f) = 0$.

Proof of Lemma 2 Let $h \in \mathcal{F}_M$ satisfy $h|_{M-N} = 0$ and $h|_p \neq 0$, then $fh|_M = 0$, and from Lemma 1 we have $v(fh) = 0$. On the other hand, the Leibniz rule also yields $v(fh) = f|_p v(h) + h|_p v(f) = h|_p v(f)$, and hence $h|_p v(f) = 0$. Note that $h|_p \neq 0$, we therefore have $v(f) = 0$. \square

Now we can prove Theorem 2.2.1. Let $f = f_1 - f_2$, then $f|_N = 0$. From Lemma 2 we know that $v(f) = 0$. On the other hand, from the linearity we know that $v(f) = v(f_1 - f_2) = v(f_1) - v(f_2)$; therefore, $v(f_1) = v(f_2)$. \square

Remark 3 Definition 2 requires that a vector v at p can only act on $f \in \mathcal{F}_M$. If a function f is only defined on a neighborhood U ($\neq M$) of $p \in M$, i.e., $f \in \mathcal{F}_U$, $f \notin \mathcal{F}_M$, then $v(f)$ is meaningless. However, one can always find an $\tilde{f} \in \mathcal{F}_M$ and a neighborhood $N \subset U$ of p such that $\tilde{f}|_N = f|_N$, and thus we can define $v(f)$ as $v(\tilde{f})$. Although for the same f there are infinitely many of \tilde{f} that satisfy the requirement above, Theorem 2.2.1 assures that $v(\tilde{f})$ are the same for all these \tilde{f} . Thus, it is legal to define $v(f)$ in terms of $v(\tilde{f})$. This conclusion is very useful. For example, suppose (O, ψ) is a coordinate system of M , then the μ th coordinate x^μ is a function on O (instead of M), but it is still valid to talk about a vector v at any point p of O acting on x^μ ; that is, $v(x^\mu)$ is meaningful.

[The End of Optional Reading 2.2.1]

According to Definition 2, to define a vector at p all we have to do is assign a map from \mathcal{F}_M to \mathbb{R} that satisfies conditions (a) and (b); that is, assign a corresponding rule according to which each $f \in \mathcal{F}_M$ corresponds to a specific real number. Since there are lots of maps like this, there are (infinitely) many vectors at p . For example, suppose (O, ψ) is a coordinate system with coordinates x^μ , then any smooth function $f \in \mathcal{F}_M$ on M combined with (O, ψ) yields a function of n variables $F(x^1, \dots, x^n)$. In this way, we can define n vectors for any point p in O , denoted by X_μ (where $\mu = 1, \dots, n$), whose action on an arbitrary $f \in \mathcal{F}_M$, i.e., $X_\mu(f)$, are defined as the following real number

$$X_\mu(f) := \left. \frac{\partial F(x^1, \dots, x^n)}{\partial x^\mu} \right|_p, \quad \forall f \in \mathcal{F}_M, \quad (2.2.1)$$

where $\partial F(x^1, \dots, x^n)/\partial x^\mu|_p$ is an abbreviation for $\partial F(x^1, \dots, x^n)/\partial x^\mu|_{(x^1(p), \dots, x^n(p))}$. We will also abbreviate $\partial F(x^1, \dots, x^n)/\partial x^\mu$ as $\partial f(x^1, \dots, x^n)/\partial x^\mu$ or $\partial f(x)/\partial x^\mu$, even $\partial f/\partial x^\mu$; the reader should recognize that the f in $\partial f/\partial x^\mu$

stands for a function of n variables $F(x^1, \dots, x^n)$ rather than a scalar field f . Thus, (2.2.1) can be shortened as

$$X_\mu(f) := \left. \frac{\partial f(x)}{\partial x^\mu} \right|_p, \quad \forall f \in \mathcal{F}_M, \quad (2.2.1')$$

Theorem 2.2.2 *Let V_p represent the collection of all vectors at p in M , then V_p is an n -dimensional vector space (where n is the dimension of M), i.e., $\dim V_p = \dim M \equiv n$.*

Proof (A) Define the addition, scalar multiplication and zero element according to the following three equations; it is not hard to verify that V_p satisfies Definition 1, and hence is a vector space.

- (1) $(v_1 + v_2)(f) := v_1(f) + v_2(f), \quad \forall f \in \mathcal{F}_M, \quad v_1, v_2 \in V_p;$
- (2) $(\alpha v)(f) := \alpha \cdot v(f), \quad \forall f \in \mathcal{F}_M, \quad v \in V_p, \quad \alpha \in \mathbb{R};$
- (3) Define the zero element $\underline{0} \in V_p$ to satisfy $\underline{0}(f) = 0, \quad \forall f \in \mathcal{F}_M.$

(B) Choose an arbitrary coordinate system whose coordinate patch contains p , then (2.2.1) defines n vectors X_μ at p , where $\mu = 1, \dots, n$. We want to show that they are linearly independent. Suppose n real numbers $\alpha^\mu (\mu = 1, \dots, n)$ are such that $\alpha^\mu X_\mu = \underline{0}$. (In this text we adopt the Einstein summation convention; that is, repeated indices are assumed to be summed over; here $\alpha^\mu X_\mu$ is short for $\sum_{\mu=1}^n \alpha^\mu X_\mu$.) Since the coordinates $x^\nu (\nu = 1, \dots, n)$ can be treated as functions on the coordinate patch, both sides of this equation should give us the same result when applied to x^μ . According to the definition of the zero element $\underline{0}$ [(3) of (A) in this proof], the action on the right-hand side yields

$$\underline{0}(x^\nu) = 0, \quad (2.2.2a)$$

while the action on the left-hand side yields

$$\alpha^\mu X_\mu(x^\nu) = \alpha^\mu \partial x^\nu / \partial x^\mu |_p = \alpha^\mu \delta^\nu_\mu = \alpha^\nu, \quad (2.2.2b)$$

where we used (2.2.1) in the first step, and δ^μ_ν is defined as $\delta^\mu_\nu = \begin{cases} 1, & \mu = \nu; \\ 0, & \mu \neq \nu. \end{cases}$. Comparing (2.2.2a) and (2.2.2b), we can see that $\alpha^\nu = 0, \nu = 1, \dots, n$. Therefore, X_1, \dots, X_n are linearly independent, and thus $\dim V_p \geq n$.

(C) To show that $\forall v \in V_p$, we have

$$v = v^\mu X_\mu, \quad (2.2.3)$$

where

$$v^\mu = v(x^\mu). \quad (2.2.3')$$

[This is the tricky step, see Wald (1984) p. 16 for a proof]. Equation (2.2.3) indicates that any element of V_p can be expressed linearly in terms of these X_μ , and (2.2.3') says that its coefficients are the real numbers given by the action of v on x^μ . The combination of (B) and (C) indicates that $\{X_1, \dots, X_n\}$ is a basis of V_p , and therefore $\dim V_p = n$. \square

Definition 3 $\{X_1, \dots, X_n\}$ of any point p in a coordinate patch is called a **coordinate basis**; each X_μ is called a **coordinate basis vector**, and the coefficients v^μ of $v \in V_p$ expressed by $\{X_\mu\}$ are called the **coordinate components** of v .

Theorem 2.2.3 Suppose $\{x^\mu\}$ and $\{x'^v\}$ are two coordinate systems, the intersection of their coordinate patches is non-empty, p is a point in the intersection, $v \in V_p$, $\{v^\mu\}$ and $\{v'^v\}$ are the coordinate components of v in these two systems, then

$$v'^v = \left. \frac{\partial x'^v}{\partial x^\mu} \right|_p v^\mu, \quad (2.2.4)$$

where x'^v is short for the coordinate transformation function $x'^v(x^\sigma)$ between these two systems.

Proof We first derive the relationship between two coordinate bases $\{X_\mu\}$ and $\{X'_v\}$ at p . From the definition of $\{X_\mu\}$ we can see that $\forall f \in \mathcal{F}_M$,

$$X_\mu(f) = \left. \frac{\partial f(x)}{\partial x^\mu} \right|_p, \quad X'_v(f) = \left. \frac{\partial f'(x')}{\partial x'^v} \right|_p,$$

where $f(x)$ and $f'(x')$ are abbreviations for $f(x^1, \dots, x^n)$ and $f'(x'^1, \dots, x'^n)$, namely the two functions of n variables coming from the combination of the scalar field $f : M \rightarrow \mathbb{R}$ and the coordinate systems $\{x^\mu\}$ and $\{x'^v\}$, respectively. Suppose q is an arbitrary point in the intersection of the two coordinate patches, then the scalar field f has the value $f|_q$ at q such that $f|_q = f(x(q)) = f'(x'(q))$, denoted by $f(x) = f'(x')$ for short. On the other hand, each x'^v corresponds to n functions of x^μ (coordinate transformation relations), denoted by $x'^v = x'^v(x)$ for short, and thus $f(x) = f'(x'(x))$. Hence,

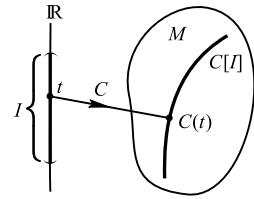
$$X_\mu(f) = \left. \frac{\partial f'(x'(x))}{\partial x^\mu} \right|_p = \left(\frac{\partial f'(x')}{\partial x'^v} \frac{\partial x'^v}{\partial x^\mu} \right)_p = \left. \frac{\partial x'^v}{\partial x^\mu} \right|_p X'_v(f), \quad \forall f \in \mathcal{F}_M.$$

The equation above indicates that the maps X_μ and $(\partial x'^v / \partial x^\mu)|_p X'_v$ are equivalent, i.e.,

$$X_\mu = \left. \frac{\partial x'^v}{\partial x^\mu} \right|_p X'_v. \quad (2.2.5)$$

Therefore, $v = v^\mu X_\mu = v'^\mu X'_\mu$ can be expressed as

Fig. 2.3 The map $C : I \rightarrow M$ is called a curve on M



$$v^\mu \frac{\partial x'^\nu}{\partial x^\mu} \Big|_p X'_\nu = v'^\nu X'_\nu .$$

Since the n basis vectors in $\{X'_\nu\}$ are linearly independent, we arrive at (2.2.4). \square

Equation (2.2.4) is called the **vector (components) transformation law**; many books use this equation as the definition of a vector.

Next, we introduce the definitions of a curve and its tangent vector.

Definition 4 Suppose I is an interval of \mathbb{R} , then a C^r function $C : I \rightarrow M$ is called a **curve** of class C^r on M . From now on, the term “curve” will refer to a smooth (C^∞) curve unless stated otherwise. For any $t \in I$, there is a corresponding unique point $C(t) \in M$ (see Fig. 2.3). t is called the **parameter** of the curve.

Remark 4 The curve described here is closely related to the intuitive concept of a curve, but there is also a difference. An intuitive curve usually refers to the image of the map $C : I \rightarrow M$ above, namely a subset $C[I]$ of M (see Fig. 2.3), without any parameter having been mentioned. The curve defined above refers to the map itself, which is a “curve with a parameter”.³ Suppose the images of the maps $C : I \rightarrow M$ and $C' : I' \rightarrow M$ coincide (see Fig. 2.4), then one would intuitively regard them as the same curve; however, as long as C and C' are different maps, according to Definition 4, they are different curves. Nevertheless, we can say in most instances that C and C' are two parametrizations of “the same curve”. To be accurate, the curve $C' : I' \rightarrow M$ is called a **reparametrization** of the curve $C : I \rightarrow M$ if \exists an onto map $\alpha : I \rightarrow I'$ satisfying (a) $C = C' \circ \alpha$, (b) the function $t' = \alpha(t)$ induced by α has a nonvanishing derivative. Here is the explanation: from $C = C' \circ \alpha$ we have

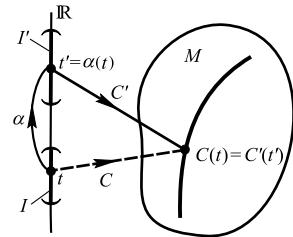
$$C(t) = C'(\alpha(t)) = C'(t') , \quad \forall t \in I .$$

The map α being onto assures $C'[I'] = C[I]$, i.e., the maps of these two curves have the same image.⁴

³ However, there also exists a curve $C : I \rightarrow M$ such that its image covers the whole M , which seems quite far from an intuitive curve.

⁴ The fact that α satisfies condition (b) assures that α has the property of one-to-one; by adding the ontoneess we can see that C is also a reparametrization of C' .

Fig. 2.4 The reparametrization of a curve



Remark 5 ① The image of a curve C is also often denoted by $C(t)$ (instead of $C[I]$) in order to indicate that the parameter of the curve is t . Note that if t is one specific value (“dead”), then $C(t)$ only stands for a point in the image of the curve; only when we consider t “can run all over I ” (“alive”) does $C(t)$ stand for the image of the curve. We also usually refer to the image of a curve simply as a curve; the reader should recognize by the context whether the word “curve” means the map or its image. ② The definition of a curve is independent of the coordinate system, and hence is absolute. However, it is convenient to express a curve explicitly with the help of a coordinate system. Suppose (O, ψ) is a coordinate system, $C[I] \subset O$, then $\psi \circ C$ is a map from $I \subset \mathbb{R}$ to \mathbb{R}^n , which amounts to n functions of one variable $x^\mu = x^\mu(t)$, $\mu = 1, \dots, n$. These n equations are called the **parametric equations**, or a **parametric representation** of the curve. A simple example is as follows: let $M = \mathbb{R}^2$, $\{x^1, x^2\}$ is the natural coordinate system of \mathbb{R}^2 , then $x^1 = \cos t$, $x^2 = \sin t$ are the parametric equations of a curve $C : \mathbb{R} \rightarrow \mathbb{R}^2$, which is a unit circle in \mathbb{R}^2 centered at the origin.

Definition 5 Suppose (O, ψ) is a coordinate system and x^μ are coordinates, then the following subset of O :

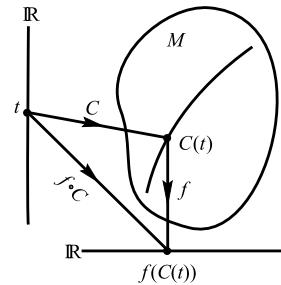
$$\{p \in O \mid x^2(p) = \text{constant}, \dots, x^\mu(p) = \text{constant}\}$$

can be regarded as (the image of) a curve with the parameter x^1 (changing the constant value of x^2, \dots, x^n gives another curve), called the **x^1 -coordinate line**. The **x^μ -coordinate line** can be defined likewise.

Example 1 In a 2-dimensional Euclidean space, the x - and y -coordinate lines of the Cartesian coordinate system $\{x, y\}$ are two sets of parallel lines that are perpendicular to each other; the φ -coordinate lines of the polar coordinate system $\{r, \varphi\}$ are an infinite number of concentric circles centered at the origin, and the r -coordinate lines are an infinite number of half-lines that start at the origin.

Now let us discuss the tangent vector of a curve. Intuitively, one may think there are infinitely many tangent vectors parallel to each other at one point of a curve. However, if we define a curve as a map (“curve with a parameter”), then there is only one tangent vector at one point of a curve. The definition is as follows:

Fig. 2.5 A function $f : M \rightarrow \mathbb{R}$ on M combined with a curve C yields a map $f \circ C : I \rightarrow \mathbb{R}$, i.e., a function of one variable $f(C(t))$



Definition 6 Suppose $C(t)$ is a C^1 curve on a manifold M , then the **tangent vector** T at $C(t_0)$ tangent to $C(t)$ is a vector at $C(t_0)$, whose action on $f \in \mathcal{F}_M$ is defined as

$$T(f) := \frac{d(f \circ C)}{dt} \Big|_{t_0}, \quad \forall f \in \mathcal{F}_M. \quad (2.2.6)$$

Remark 6 ① $f : M \rightarrow \mathbb{R}$ is a function (scalar field) on M but not generally a function of one variable. However, $f \circ C$, the combination of f and a curve $C : I \rightarrow M$, is a function of one variable with the argument t [also denoted by $f(C(t))$, see Fig. 2.5]. When there is no confusion, $d(f \circ C)/dt$ can also be denoted by df/dt . ② The tangent vector at a point $C(t_0)$ tangent to $C(t)$ is also often denoted by $\partial/\partial t|_{C(t_0)}$, and thus (2.2.6) can also be written as

$$\frac{\partial}{\partial t} \Big|_{C(t_0)} (f) := \frac{d(f \circ C)}{dt} \Big|_{t_0} = \frac{df(C(t))}{dt} \Big|_{t_0}, \quad \forall f \in \mathcal{F}_M. \quad (2.2.6')$$

Example 2 Since the x^μ -coordinate line is a curve with x^μ as the parameter, the coordinate basis vector X_μ at p defined in (2.2.1) is a tangent vector of the x^μ -coordinate line passing through p . Hence, it is also usually denoted by $\partial/\partial x^\mu|_p$, and therefore (2.2.1') can also be expressed as

$$\frac{\partial}{\partial x^\mu} \Big|_p (f) := \frac{\partial f(x)}{\partial x^\mu} \Big|_p \quad \forall f \in \mathcal{F}_M. \quad (2.2.6'')$$

Thus, the symbol $\partial f/\partial x^\mu$ can be interpreted as either $\partial F(x^1, \dots, x^n)/\partial x^\mu$ [see (2.2.1)], or the action of a coordinate line tangent vector $\partial/\partial x^\mu$ on a scalar field f .

Theorem 2.2.4 Suppose the parametric equations of a curve $C(t)$ in a given coordinate system is $x^\mu = x^\mu(t)$, then the expansion of the tangent vector at an arbitrary point on the curve in this coordinate basis gives

$$\frac{\partial}{\partial t} = \frac{dx^\mu(t)}{dt} \frac{\partial}{\partial x^\mu}. \quad (2.2.7)$$

That is, the coordinate components of the tangent vector $\partial/\partial t$ of the curve $C(t)$ is the derivative of the parametric representation $x^\mu(t)$ of $C(t)$ in this system with respect to t .

Proof Exercise. □

Definition 7 Two nonzero vectors $v, u \in V_p$ are said to be **parallel** if $\exists \alpha \in \mathbb{R}$ such that $v = \alpha u$.

From Definition 6 we can see that the tangent vectors of a curve depend on the parametrization of the curve; there is only one vector at each point $C(t_0)$ of a curve $C(t)$ that is tangent to $C(t)$. The reason why it intuitively seems that there are infinitely many (parallel) tangent vectors at one point of a curve is that, in that case, we understand a curve as being the image of the map rather than the map itself [making “degenerate” an infinite number of curves (maps) with the same image into one curve]. The theorem below indicates that if two curves C and C' have the same image, then their tangent vectors at any point are parallel.

Theorem 2.2.5 Suppose a curve $C' : I' \rightarrow M$ is a reparametrization of $C : I \rightarrow M$, then their tangent vectors at any image point has the following relation:

$$\frac{\partial}{\partial t} = \frac{dt'(t)}{dt} \frac{\partial}{\partial t'}, \quad (2.2.8)$$

where $t'(t)$ is the function of one variable induced by a map $\alpha : I \rightarrow I'$ (see Remark 4), i.e., $\alpha(t)$.

Proof We can combine any $f \in \mathcal{F}_M$ with C and C' to induce functions $f(C(t))$ and $f(C'(t'))$ of one variable, denoted by $f(t)$ and $f'(t')$, respectively. Suppose the image of $t \in I$ under the map $(C'^{-1} \circ C)$ is t' , then $f(t) = f'(t'(t))$ (see Fig. 2.6). Hence,

$$\frac{\partial}{\partial t}(f) = \frac{df(t)}{dt} = \frac{df'(t'(t))}{dt} = \frac{df'(t')}{dt'} \frac{dt'}{dt} = \frac{\partial}{\partial t'}(f) \frac{dt'}{dt} = \left(\frac{dt'}{dt} \frac{\partial}{\partial t'} \right)(f), \quad \forall f \in \mathcal{F}_M,$$

and thus we have (2.2.8). □

From Definition 6 we can see that, $\forall p \in M$, if we choose an arbitrary curve $C(t)$ such that $p = C(t_0)$, then there must be an element in V_p that can be regarded as the tangent vector of this curve at $C(t_0)$. Now we ask: if we choose an arbitrary element v in V_p , can we find a curve that passes through p whose tangent vector at p is v ? The answer is affirmative: this kind of curve not only exists, but also they are numerous (Fig. 2.7 is a visual representation). For instance, if we choose an arbitrary coordinate system $\{x^\mu\}$ such that p is contained in its coordinate patch, then the curve with the parametric equations $x^\mu(t) = x^\mu|_p + v^\mu t$ is such a curve, where v^μ is the coordinate components of v in this system.

In conclusion, any element in V_p can be viewed as the tangent vector of a curve passing through p . Therefore, a vector at p is also called a **tangent vector**, and V_p is called the **tangent space** at p .

Fig. 2.6 Figure for the proof of Theorem 2.2.5

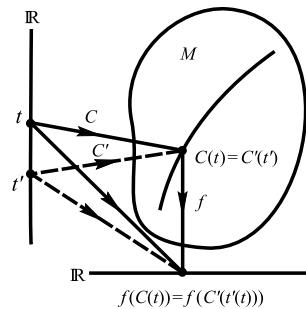
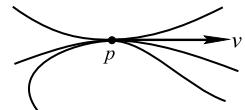


Fig. 2.7 A vector v at p is the common tangent vector of many curves



2.2.2 Tangent Vector Fields on Manifolds

Definition 8 Suppose A is a subset of M . If we assign a vector to each point of A , we obtain a **vector field** defined on A .

Example 3 The tangent vectors at all points of a non-self-intersecting curve $C(t)$ form a vector field on $C(t)$ (as a subset of M).

Suppose v is a vector field on M and f is a function on M , then the value of v at any point p of M will map f to a real number $v|_p(f)$ according to Definition 2, which forms a function $v(f)$ when varying p over M . Therefore, a vector field v can be viewed as a map that turns a function f into a function $v(f)$.

Definition 9 A vector field v on M is said to be of **class C^∞ (smooth)** if the result of v acting on a C^∞ function is a C^∞ function, i.e., $v(f) \in \mathcal{F}_M$, $\forall f \in \mathcal{F}_M$. v is said to be of **class C^r** if the result of v acting on a C^∞ function is a C^r function.

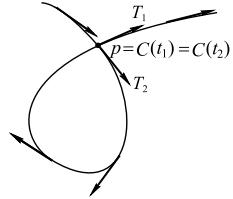
From now on, unless stated otherwise, the term “vector field” will mean smooth (C^∞) vector field.

Example 4 (1) A set of coordinate basis vectors $\{X_\mu \equiv \partial/\partial x^\mu\}$ form n smooth vector fields in the coordinate patch, called **coordinate basis vector fields**. (2) The electric field \vec{E} of a point charge at point q in \mathbb{R}^3 is a smooth vector field on the manifold $M \equiv \mathbb{R}^3 - \{q\}$.

[Optional Reading 2.2.2]

If $C : I \rightarrow M$ is a self-intersecting curve, i.e., $\exists t_1, t_2 \in I$ such that $C(t_1) = C(t_2) \equiv p \in M$, then there are two tangent vectors at p , and thus we cannot say the tangent vectors form a vector field defined on $C(t)$ (the image of map C). However, we can define a **vector field along a curve C** (the map C) [see Spivak (1970) Vol. II, p. 247; Sachs and Wu (1977)]

Fig. 2.8 A point p on a self-intersecting curve $C[I]$ has two tangent vectors T_1 and T_2 , but one can still talk about the tangent vector field along the curve (map) C



pp. 36–37], which is a map that corresponds each $t \in I$ to a $v \in V_{C(t)}$ whose domain is I rather than the subset $C[I]$ of M . Thus, this vector field can be denoted by $v(t)$. In the case of Fig. 2.8, both “the tangent vector field on $C[I]$ ” and “the tangent vector of $C[I]$ at p ” are meaningless, but “the tangent vector field $T(t)$ along C ” is meaningful. Also, we can talk about the tangent vector $T(t_1)$ of C at t_1 (T_1 in the figure) and the tangent vector $T(t_2)$ of C at t_2 (T_2 in the figure). In this text, we often generally use “a vector field on a curve $C(t)$ ”; for a self-intersecting curve, this actually means the vector field along C .

[The End of Optional Reading 2.2.2]

Theorem 2.2.6 *A necessary and sufficient condition for a vector field v on M to be C^∞ (or C^r) is that its components in any coordinate basis are C^∞ (or C^r) functions.*

Proof Exercise. □

Suppose v is a smooth vector field on M , then $v(f) \in \mathcal{F}_M, \forall f \in \mathcal{F}_M$. If u is another smooth vector field on M , then $u(v(f)) \in \mathcal{F}_M$. However, the function $u(v(f)) \in \mathcal{F}_M$ is not necessarily equal to $v(u(f)) \in \mathcal{F}_M$, and thus we have the following definition:

Definition 10 The **commutator** of two smooth vector fields u and v is a smooth vector field $[u, v]$, defined as

$$[u, v](f) := u(v(f)) - v(u(f)), \quad \forall f \in \mathcal{F}_M. \quad (2.2.9)$$

Remark 7 The equation above is the definition of the commutator $[u, v]$ (as a vector field), the definition of its value $[u, v]|_p$ at each point should be understood as

$$[u, v]|_p(f) := u|_p(v(f)) - v|_p(u(f)) \quad \forall f \in \mathcal{F}_M. \quad (2.2.9')$$

To firmly believe that $[u, v]|_p$ defined by the equation above is a vector at p , one should also show that (Exercise 2.8) it satisfies the two conditions in Definition 2.

Theorem 2.2.7 Suppose $\{x^\mu\}$ is an arbitrary coordinate system, then $[\partial/\partial x^\mu, \partial/\partial x^\nu] = 0, \mu, \nu = 1, \dots, n$.

Proof Suppose $f(x)$ is the function of n variables that comes from the combination of f and this coordinate system. From calculus we know that

$$\frac{\partial}{\partial x^\mu} \frac{\partial}{\partial x^\nu} f(x) = \frac{\partial}{\partial x^\nu} \frac{\partial}{\partial x^\mu} f(x)$$

and the theorem can be proved instantly. \square

Theorem 2.2.7 indicates that two arbitrary basis vector fields of any coordinate system commute with each other.⁵

Definition 11 A curve $C(t)$ is called an **integral curve** of a vector field v if the tangent vector at each point on it equals the value of v at this point.

Theorem 2.2.8 Suppose v is a smooth vector field on M , then for any point p of M there exists a unique integral curve $C(t)$ of v passing through it [which satisfies $C(0) = p$] (“unique” should be understood as “locally unique”, see Optional Reading 2.2.3).

Proof Choose an arbitrary coordinate system $\{x^\mu\}$ whose coordinate patch contains p . Suppose the parametric equations of the integral curve is $x^\mu = x^\mu(t)$, then it follows from (2.2.7) that $x^\mu(t)$ satisfies the first-order ordinary differential equations

$$\frac{dx^\mu(t)}{dt} = v^\mu(x^1(t), \dots, x^n(t)), \quad \mu = 1, \dots, n,$$

where v^μ is the μ th component of v in this coordinate basis field, which is a given function of x^1, \dots, x^n . From calculus we know that there exists a unique solution for this system of equations under given initial conditions $x^\mu(0)$ ($\mu = 1, \dots, n$). When a point p is given, a set of initial conditions is given, namely $x^\mu(0) = x^\mu|_p$; hence, there must be a unique solution $x^1(t), \dots, x^n(t)$, and the curve satisfied by these n equations is the integral curve we want. It should also be verified that the curve obtained in this way is independent of the coordinate system; the proof is left to the reader. \square

[Optional Reading 2.2.3]

The word “unique” in Theorem 2.2.8 should be understood as “locally unique”. Suppose you have found an integral curve $C : (a, b) \rightarrow M$ of v , and $0 \in (a, b)$, $C(0) = p$. Your friend can always choose a smaller interval $(a', b') \subset (a, b)$ that contains 0 and define a new curve $C' : (a', b') \rightarrow M$ as $C'(t) = C(t) \forall t \in (a', b')$. The domains of the map C' and C are not equivalent, and hence $C' \neq C$. In this sense, the integral curves which pass through p are not unique. However, C is nothing but an extension of C' ; they are locally the same, and Theorem 2.2.8 holds as long as we interpret “unique” as “locally unique”.

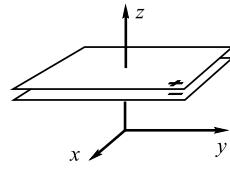
Speaking of extension, one may naturally ask: Is it always possible to extend the domain of an integral curve C from (a, b) to the entirety of \mathbb{R} ? The answer is negative. The following simple example is helpful for understanding this. Let x_1 and x_2 be the natural coordinates of \mathbb{R}^2 , define a curve $C : \mathbb{R} \rightarrow \mathbb{R}^2$ as $C(t) := (0, t) \in \mathbb{R}^2 \forall t \in \mathbb{R}$. The image of this curve is

⁵ Conversely, suppose X_1, \dots, X_n are n C^∞ vector fields on a neighborhood N of $p \in M$ that are linearly independent everywhere, and

$$[X_\mu, X_\nu] = 0, \quad \mu, \nu = 1, \dots, n,$$

then there must exist a coordinate system $\{x^\mu\}$ whose coordinate patch $O \subset N$ contains p , and on O we have $X_\mu = \partial/\partial x^\mu$, $\mu = 1, \dots, n$. See Wald (1984) Exercise 5 in Chap. 2 for a hint of the proof of this theorem. For a complete proof see Spivak (1970) Vol. I, pp. 219–220.

Fig. 2.9 The charge distribution of a parallel-plate capacitor has translational symmetry along the x - and y -axes



the x^2 -coordinate axis in \mathbb{R}^2 . It is not difficult to see that this is the integral curve of the vector field $\partial/\partial x^2$ passing through $(0, 0)$. If we cut out the “upper half” of \mathbb{R}^2 and regard the rest of it as a manifold M , or more precisely, define M as $M := \{(x^1, x^2) \in \mathbb{R}^2 | x^2 < 1\}$, then the map C has no image at $t \geq 1$, and its domain is just the open interval $(-\infty, 1)$ instead of \mathbb{R} . This domain cannot be extended anymore, and thus the map $C : (-\infty, 1) \rightarrow M$ is called an inextensible integral curve of the vector field $\partial/\partial x^2$ on M . Therefore, Theorem 2.2.8 can also be expressed as:

Theorem 2.2.8' Suppose v is a smooth (actually, C^1 is already enough) vector field on M , then for any point p in M there exists a unique inextensible integral curve $C(t)$ passing through it [and satisfying $C(0) = p$].

[The End of Optional Reading 2.2.3]

We will use some basic knowledge of group theory below, so we introduce the following definition as a supplement (see Appendix G in Volume II for a detailed introduction of the theory of Lie groups and Lie algebras):

Definition 12 A **group** is a set G together with a map $G \times G \rightarrow G$ (called the **group multiplication**, the product of elements g_1 and g_2 is denoted by $g_1 g_2$), that satisfies the following conditions:

- (a) $(g_1 g_2) g_3 = g_1 (g_2 g_3)$, $\forall g_1, g_2, g_3 \in G$;
- (b) \exists an **identity element** e such that $eg = ge = g$, $\forall g \in G$;
- (c) $\forall g \in G$, \exists an **inverse element** $g^{-1} \in G$ such that $g^{-1}g = gg^{-1} = e$.

Symmetry has a great significance for physics, and group theory is a powerful tool for the study of symmetry. If an object is invariant under a certain transformation, then we say it has a symmetry under this transformation. Take Fig. 2.9 as an example; consider a moving point on a charged plane that is translating along the x - (or y -) axis. Since the surface charge density σ at the moving point is invariant under translation, we say σ has translational symmetry along the x - (or y -) axis. More precisely, the translational symmetry of σ along the x -axis means that the function $\sigma(x, y, z)$ satisfies

$$\sigma(x, y, z) = \sigma(x + a, y, z), \quad \forall a \in \mathbb{R}, \quad (2.2.10)$$

where the point transformation represented by

$$x \mapsto x + a, \quad y \mapsto y, \quad z \mapsto z \quad (2.2.11)$$

is called a **translation** along the x -axis. Suppose G is the collection of all the translations along the x -axis, then an element in G can be characterized by a real number a , denoted by $\phi_a \in G$. Consider $p \equiv (x, y, z)$ and $q \equiv (x + a, y, z)$ as two points in \mathbb{R}^3 , then the transformation (2.2.11) corresponds to the map $\phi_a : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ [satisfying $\phi_a(p) = q$], which is a diffeomorphism. Moreover, if we define the multiplication for G as

$$\phi_a \phi_b := \phi_{a+b}, \quad \forall \phi_a, \phi_b \in G, \quad (2.2.12)$$

then G forms a group (ϕ_0 is the identity, and ϕ_{-a} is the inverse of ϕ_a). Each of the infinitely many elements of this group can be characterized by a real number a , which is therefore called a **parameter**, and G is called a **one-parameter group**. Also, since each group element $\phi_a \in G$ is a diffeomorphism on \mathbb{R}^3 , we also call G a **one-parameter group of diffeomorphisms** on \mathbb{R}^3 . To help the reader understand the definition of a one-parameter group of diffeomorphisms, let us set the stage for it first. Suppose M is a manifold, then $\mathbb{R} \times M$ is a manifold that has one more dimension than M (see the last paragraph of Sect. 2.1). Suppose ϕ is a map from $\mathbb{R} \times M$ to M (i.e., $\phi : \mathbb{R} \times M \rightarrow M$), then it can turn a real number $t \in \mathbb{R}$ and a point $p \in M$ into a point $\phi(t, p) \in M$. We can also visualize ϕ as a machine with two slots, denoted by $\phi(\bullet, \bullet)$; in order to produce an “end product” $\phi(t, p) \in M$, one has to input two “raw materials”, namely, a real number $t \in \mathbb{R}$ and a point $p \in M$. If we input $t \in \mathbb{R}$ alone, then what it can produce is only a semi-manufacture $\phi(t, \bullet)$, which is also a machine that gives an “end product” after we input $p \in M$. $\phi(t, \bullet)$ is usually denoted by ϕ_t , i.e., $\phi_t : M \rightarrow M$. On the other hand, if we input $p \in M$ to $\phi(\bullet, \bullet)$ first, we get a semi-manufacture $\phi(\bullet, p)$, which is also a machine waiting for $t \in \mathbb{R}$ to be input. $\phi(\bullet, p)$ is usually denoted by ϕ_p , i.e., $\phi_p : \mathbb{R} \rightarrow M$.

Definition 13 A C^∞ map $\phi : \mathbb{R} \times M \rightarrow M$ is called a **one-parameter group of diffeomorphisms** on M if

- (a) $\phi_t : M \rightarrow M$ is a diffeomorphism $\forall t \in \mathbb{R}$;
- (b) $\phi_t \circ \phi_s = \phi_{t+s}$, $\forall t, s \in \mathbb{R}$.

Remark 8 A set $\{\phi_t | t \in \mathbb{R}\}$ is a group with map composition as the multiplication, whose group elements ϕ_t are diffeomorphisms from M to M , and ϕ_0 is the identity. [from Definition 13(b) we know that $\phi_t \circ \phi_0 = \phi_t$, and hence ϕ_0 is the identity map.] “ $\phi : \mathbb{R} \times M \rightarrow M$ is a one-parameter group of diffeomorphisms on M ” actually means that $\{\phi_t | t \in \mathbb{R}\}$ is a one-parameter group of diffeomorphisms.

Suppose $\phi : \mathbb{R} \times M \rightarrow M$ is a one-parameter group of diffeomorphisms, then $\forall p \in M$, $\phi_p : \mathbb{R} \rightarrow M$ is a smooth curve that passes through p [satisfying $\phi_p(0)=p$], called the **orbit** of this one-parameter group of diffeomorphisms that passes through p . Denote the tangent vector of this curve at $\phi_p(0)$ by $v|_p$, then we have a smooth vector field v on M . Thus, a one-parameter group of diffeomorphisms on M gives rise to a smooth vector field on M . Now let us see if the converse holds or not. Suppose v is a smooth vector field on M , it seems that $\forall t \in \mathbb{R}$ one can use its integral curve to define a diffeomorphism ϕ_t from M to M . [$\forall p \in M$, define $\phi_t(p)$ as the point such

that it is located on the integral curve that passes through p , the difference of whose parameter and the parameter of p is t .] So it looks like we can obtain a one-parameter group of diffeomorphisms. However, the following problem may occur: the image point does not exist for certain parameters of a curve (cutting out a region M could make this situation happen); therefore, we can only say that a smooth vector field on M gives rise to a one-parameter local group of diffeomorphisms, see Optional Reading 2.2.4.

[Optional Reading 2.2.4]

Suppose the integral curve C of a vector field v that passes through p has a range of parameters that cannot reach all of \mathbb{R} (see the second paragraph of Optional Reading 2.2.3), namely $\exists t \in \mathbb{R}$ such that $C(t)$ is not a point of M , then ϕ_t defined above is not even a map from M to M [at least the image point $\phi_t(p)$ does not exist], so clearly not a diffeomorphism from M to M . However, it can be proved that $\forall p_0 \in M$, one can always find an open neighborhood U of p_0 and an open interval I in \mathbb{R} that contains 0 to make the map ϕ in the text above meaningful when restricted on $I \times U$ (i.e., there exists a map $\phi : I \times U \rightarrow M$). The precise definition is $\forall t \in I$, $\phi_t : U \rightarrow M$ is a map that maps any $p \in U$ to a point on the integral curve that passes through p , with t the difference between the parameters of p and this point (the reader may understand this with the help of the simple example in the second paragraph of Optional Reading 2.2.3). Moreover, it can also be proved that $\phi : I \times U \rightarrow M$ has the following properties:

- (a) $\forall t \in I$, $\phi_t : U \rightarrow \phi_t[U]$ is a diffeomorphism;
- (b) If $t, s, t + s \in I$ (real numbers t, s and $t + s$ are all in the open interval I), then $\phi_t \circ \phi_s = \phi_{t+s}$.

Such a $\{\phi_t \mid t \in I\}$ is called a **one-parameter local group of diffeomorphisms** or a **one-parameter family of diffeomorphisms**.

A vector field is said to be **complete** if the range of the parameter of every (inextensible) integral curve is \mathbb{R} . Obviously, each complete smooth vector field can produce a one-parameter group of diffeomorphisms. It can be proved that any vector field on a compact manifold is complete. [References for this optional reading: Hawking and Ellis (1973) p. 27; Straumann (1984) pp. 21–22.]

[The End of Optional Reading 2.2.4]

2.3 Dual Vector Fields

Definition 1 Suppose V is a finite dimensional vector space on \mathbb{R} . A linear map $\omega : V \rightarrow \mathbb{R}$ is called a **dual vector** on V . The collection of all the dual vectors on V is called the **dual vector space** of V , denoted by V^* .⁶

Remark 1 Since addition and scalar multiplication are each defined on V , the linearity requirement for ω can be explicitly written as

$$\omega(\alpha v + \beta u) = \alpha\omega(v) + \beta\omega(u), \quad \forall v, u \in V, \quad \alpha, \beta \in \mathbb{R}. \quad (2.3.1)$$

⁶ When talking about dual vectors (and tensors, see Sect. 2.4) in the future, unless stated otherwise, we will always assume V is a finite dimensional vector space on \mathbb{R} .

Example 1 Suppose V is the collection of all 2×1 real matrices, then it forms a 2-dimensional vector space under the rule of matrix addition and scalar multiplication. Let ω represent an arbitrary 1×2 real matrix (c, d) , whose action on any element $v = \begin{pmatrix} a \\ b \end{pmatrix}$ in V can be defined using matrix multiplication: $\omega(v) := (c, d) \begin{pmatrix} a \\ b \end{pmatrix} = (ac + bd)$. The result is a 1×1 real matrix, which can be identified as a real number $ac + bd$. A map $\omega : V \rightarrow \mathbb{R}$ defined in this way is clearly linear, and thus any 1×2 real matrix is a dual vector on V . More generally, if we consider column matrices ($n \times 1$ matrices) as vectors, then row matrices ($1 \times n$ matrices) are dual vectors.

Theorem 2.3.1 V^* is a vector space, and $\dim V^* = \dim V$.

Proof Define addition, scalar multiplication and the zero element for V^* as follows:

$$\begin{aligned} (\omega_1 + \omega_2)(v) &:= \omega_1(v) + \omega_2(v), \quad \forall \omega_1, \omega_2 \in V^*, v \in V; \\ (\alpha\omega)(v) &:= \alpha \cdot \omega(v), \quad \forall \omega \in V^*, v \in V, \alpha \in \mathbb{R}; \\ \underline{0}(v) &:= 0 \in \mathbb{R}, \quad \forall v \in V. \end{aligned}$$

It is not difficult to see that such a V^* is a vector space. Suppose $\{e_\mu\}$ is a basis of V , we can define n special elements e^{1*}, \dots, e^{n*} in V^* using the following equation:

$$e^{\mu*}(e_v) := \delta^\mu{}_v, \quad \mu, v = 1, \dots, n. \quad (2.3.2)$$

The equation above only defines the action of $e^{\mu*}$ on the basis vectors in V , but since the action of $e^{\mu*}$ is linear, it actually defines the action of $e^{\mu*}$ on an arbitrary element in V . Now we only have to show that $\{e^{\mu*}\}$ is a basis of V^* . It is easy to show that e^{1*}, \dots, e^{n*} are linearly independent to each other (exercise). $\forall \omega \in V^*$, let

$$\omega_\mu \equiv \omega(e_\mu), \quad \mu = 1, \dots, n, \quad (2.3.3)$$

and then one can easily show that (Exercise 2.11)

$$\omega = \omega_\mu e^{\mu*}. \quad (2.3.4)$$

(Hint: the equation above is an equality of dual vectors, note that the action of ω on v is linear, all we have to do for proving this equation is to verify that both sides of it acting on any basis vector e_v give the same real number.) Equation (2.3.4) indicates that any element in V^* can be expressed linearly in terms of $\{e^{\mu*}\}$, and thus $\{e^{\mu*}\}$ is a basis of V^* , called the **dual basis** to the basis $\{e_\mu\}$, from which we find that $\dim V^* = \dim V$. \square

Review. Two vector spaces are said to be **isomorphic** if there exists a one-to-one and onto linear map between them (this map is called an **isomorphism**). A necessary and sufficient condition of two vector spaces to be isomorphic is they have the same dimension.

Since $\dim V^* = \dim V$, of course V^* is isomorphic to V . An isomorphism is not difficult to find. For example, suppose $\{e_\mu\}$ is a basis of V , and $\{e'^{\mu*}\}$ is the dual basis of it, then the linear map defined by $e_\mu \rightarrow e'^{\mu*}$ is an isomorphism. However, the choice of $\{e_\mu\}$ is quite arbitrary, and the isomorphism defined in this manner will change when the basis is changed; as a matter of fact, there does not exist a special (distinguishing) isomorphism between V and V^* unless an additional structure is added to V (see Sect. 2.5).

Since V^* is a vector space, it naturally has a dual space, denoted by V^{**} . Unlike the relationship between V and V^* , there exists a natural, distinguishing isomorphism that is defined as follows: $\forall v \in V$, we want a naturally defined image $v^{**} \in V^{**}$. Since V^{**} is the dual space of V^* , v^{**} should be a linear map from V^* to \mathbb{R} . Giving it a definition is nothing but establishing a rule, according to which every $\omega \in V^*$ corresponds to a unique real number $v^{**}(\omega)$. Since v^{**} is about to be defined as the image of v , $v^{**}(\omega)$ should relate to both v and ω ; seeing that the simplest real number one can construct from v and ω is $\omega(v)$, it is natural to define v^{**} as

$$v^{**}(\omega) := \omega(v) \quad \forall \omega \in V^*. \quad (2.3.5)$$

This map $V \rightarrow V^{**}$ is an isomorphism (the proof is left as Exercise 2.13). This natural isomorphic relation indicates that V and V^{**} can be viewed as the same space (identify each $v \in V$ with its image $v^{**} \in V^{**}$). Therefore, it is V and V^* that are actually useful; one cannot get any more useful spaces from extra dualities no matter how many times it is applied [for the precise meaning of natural isomorphism see Spivak (1970) Chap. 4, Exercise 6].

Theorem 2.3.2 *If there is a basis transformation $e'_\mu = A^\nu{}_\mu e_\nu$ in a vector space V ($A^\nu{}_\mu$ is simply the v th component of a new basis vector e'_μ expanded by the old basis), and the (non-degenerate) matrix constituted by elements $A^\nu{}_\mu$ is denoted by A , then the corresponding dual basis transformation is*

$$e'^{\mu*} = (\tilde{A}^{-1})_\nu{}^\mu e^{\nu*}, \quad (2.3.6)$$

where \tilde{A} is the transposed matrix of A , and \tilde{A}^{-1} is the inverse of \tilde{A} .

Remark 2 The reader may be used to writing matrix elements as $A_{v\mu}$, here we write them as $A^\nu{}_\mu$. The reason for distinguishing the upper and lower indices is to make the summation crystal clear (an upper v together with a lower v implies the summation over v) and to distinguish the type of a tensor (see Sect. 2.4 for details). However, what is important in the matrix operation is just differentiating the left and right indices. Therefore, if you want, you may change all the upper indices to lower indices for now; for instance, (2.3.6) may be written as $e'_\mu{}^* = (\tilde{A}^{-1})_{v\mu} e_v{}^*$.

Proof All we have to prove is that both sides of the equation give the same result when applied to e'_α . The proof is as follows:

$$\begin{aligned} (\tilde{A}^{-1})_v^\mu e^{v*}(e'_\alpha) &= (\tilde{A}^{-1})_v^\mu e^{v*}(A^\beta_\alpha e_\beta) = A^\beta_\alpha (\tilde{A}^{-1})_v^\mu e^{v*}(e_\beta) \\ &= \tilde{A}_\alpha^\beta (\tilde{A}^{-1})_v^\mu \delta^\nu_\beta = \tilde{A}_\alpha^\nu (\tilde{A}^{-1})_v^\mu = \delta_\alpha^\mu = e'^{\mu*}(e'_\alpha), \end{aligned}$$

where we used the linearity of the action by a dual vector on a vector in the second equality, the definitions of a transpose matrix and inverse matrix in the third and fifth equality respectively, and the definition of dual vector basis (2.3.2) in the sixth equality. \square

The discussions above all pertain to algebras; now we get back to a manifold M . Since $p \in M$ has a vector space V_p , it also has a V_p^* . If we assign a dual vector at each point of M (or $A \subset M$), we obtain a **dual vector field** on M (or A). A dual vector field ω on M is said to be **smooth** if $\omega(v) \in \mathcal{F}_M \forall$ smooth vector fields v .

Suppose $f \in \mathcal{F}_M$, let us show that f naturally induces a dual vector field on M , denoted by df . (The df that our readers are familiar with stands for the differential of a function f . From the perspective of differential geometry, the differential of f is essentially a dual vector field. Optional Reading 2.3.1 will introduce the connection between this brand new understanding and classical calculus.) To define df we only have to give the definition of its value $df|_p \in V_p^*$ at any point p of M , and to define $df|_p$ we only have to specify the real number that comes from its action on an arbitrary vector $v \in V_p$ at p . This number should be related to both f and v , and the most natural (simplest) real number that can be constructed from f and v is $v(f)$; therefore, we define $df|_p$ as

$$df|_p(v) := v(f), \quad \forall v \in V_p. \quad (2.3.7)$$

From this it is easy to show that

$$d(fg)|_p = f|_p(dg)|_p + g|_p(df)|_p, \quad (2.3.8)$$

which is exactly the Leibniz rule satisfied by the differential operator d .

Suppose (O, ψ) is a coordinate system, then the μ th coordinate x^μ can be viewed as a function on O , and thus dx^μ (as a special df) is a dual vector field defined on O . Suppose $p \in O$ and $\partial/\partial x^\nu$ is the ν th coordinate basis vector of V_p , then from (2.3.7) we know that at p we have

$$dx^\mu \left(\frac{\partial}{\partial x^\nu} \right) = \frac{\partial}{\partial x^\nu}(x^\mu) = \delta^\mu_\nu.$$

Comparing this with (2.3.2) we can see that $\{dx^\mu|_p\}$ is exactly the **dual coordinate basis** that corresponds to the coordinate basis $\{\partial/\partial x^\nu|_p\}$. The equation above holds at any point of O . Therefore, just like $\partial/\partial x^\nu$ is the ν th coordinate basis vector field on O , dx^μ is the μ th dual coordinate basis vector field on O , and $\{dx^\mu\}$ is a dual coordinate basis field on O . Any dual vector field ω on O can be expanded in terms of $\{dx^\mu\}$:

$$\omega = \omega_\mu dx^\mu, \quad (2.3.9)$$

where ω_μ are called the coordinate components of ω in this coordinate system whose expression can be obtained from (2.3.3) as

$$\omega_\mu = \omega(\partial/\partial x^\mu). \quad (2.3.10)$$

Theorem 2.3.3 Suppose (O, ψ) is a coordinate system, f is a smooth function on O , and $f(x)$ denotes the function of n variables $f(x^1, \dots, x^n)$ corresponding to $f \circ \psi^{-1}$, then df can be expanded using a dual basis $\{dx^\mu\}$ as follows:

$$df = \frac{\partial f(x)}{\partial x^\mu} dx^\mu, \quad \forall f \in \mathcal{F}_O. \quad (2.3.11)$$

Proof All we have to prove is that we obtain the same result after applying both sides of this equation to any coordinate basis vector $\partial/\partial x^\nu$, which is very straightforward. \square

Theorem 2.3.4 Suppose the coordinate patches of the coordinate systems $\{x^\mu\}$ and $\{x'^\nu\}$ have an intersection, and ω is a dual vector at an arbitrary point p in the intersection, then the transformation relation between ω_μ and ω'_ν , the components of ω in these two coordinate systems, is

$$\omega'_\nu = \left. \frac{\partial x^\mu}{\partial x'^\nu} \right|_p \omega_\mu. \quad (2.3.12)$$

Proof Exercise 2.12. \square

[Optional Reading 2.3.1]

Now we will develop an understanding of df . First, let us talk about classical calculus. Consider a function $y = f(x)$. Suppose that when the argument x is shifted by a small increment Δx at x_0 , and the corresponding shift of the function y is Δy . If $f(x)$ is a linear function, i.e., $y = ax + b$ (a, b are constants), then $\Delta y = a\Delta x$, i.e., Δy is proportional to Δx . If $f(x)$ is a nonlinear function, then $\Delta y = a\Delta x + \varepsilon$, where $a \equiv f'(x_0)$, $\varepsilon \neq 0$. In classical calculus $a\Delta x$ is called the differential of $y = f(x)$, denoted by dy , and it may be proved that ε is a higher order infinitesimal than Δx when Δx approaches zero. Δx may also be denoted by dx , and thus $dy = f'(x_0)dx$. However, an “infinitesimal nonzero quantity” is a very subtle concept that involves logical inconsistencies. Gottfried Leibniz was criticized by many mathematicians of his time when he introduced and used this concept [see Kline (1980) for details], and still to this day there are mathematicians who do not prescribe to it. [For example, see Spivak (1970) Vol. I. This book points out on p. 153 that an “infinitely small” change dx^i is “nonsense”.] We do not mean to discuss the validity of these subtle issues; but rather, we only want to explain that modern differential geometry has already provided df , the differential of a function, a brand new interpretation that is independent of the concept of “infinitesimal nonzero quantities”: df is a clearly defined dual vector field. Suppose $\{x^\mu\}$ is a coordinate system of a manifold M with a coordinate patch O , and f is a function on M , i.e., $f : M \rightarrow \mathbb{R}$, then f induces a function $f(x^1, \dots, x^n)$ of n variables. Suppose $p \in O$; classical calculus attempts to describe $df|_p$ as an (infinitesimal) increment of the function value at p . However, this increment is not yet certain since it depends on “how far along which direction” a moving point would “move” from p . Now that a vector

v at p shows exactly “how far along which direction” a moving point would potentially “move” from p , we can let $df|_p$ actually “become” a real number (increment) by assigning a $v \in V_p$. And since $df|_p$ evaluates to a real number when v is given, $df|_p$ is actually a map from the tangent space V_p of p to \mathbb{R} . To ensure that $df|_p$ has the properties of differential from classical calculus, this map is also required to be linear. Thus, $df|_p$ is a dual vector on V_p while df is a dual vector field on O . This is the most concrete and precise interpretation of df .

Physicists usually do not make any distinction between df and Δf , and like to say “ $df|_p$ equals $f(q) - f(p)$, where q is a point infinitely close to p . ” They may even sketch two points p and q on paper. In fact, p and q can not be infinitely close as long as they have been assigned (marked out in a picture), which means $f(q) - f(p)$ is not an infinitesimal quantity, and hence it can only be Δf instead of df . However, since certain approximations are always allowed in physics, treating Δf as being small enough that it approximates df is not only allowed, but often quite useful. In fact, suppose a curve $C(t)$ satisfies $C(0) = p$, $(\partial/\partial t)|_p = v$, and $q = C(\alpha)$ with α small enough, then from (2.3.7) and (2.2.6') we can see that the result of $df|_p$ acting on αv is

$$\begin{aligned} df|_p(\alpha v) &= \alpha v(f) = \alpha \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \{f[C(\Delta t)] - f[C(0)]\} \\ &\cong \alpha \frac{1}{\alpha} [f(q) - f(p)] = f(q) - f(p) \equiv \Delta f, \end{aligned}$$

and we see that (after acting on αv) $df|_p$ really gives us Δf approximately. Albert Einstein once said: “As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.” As a physics book, this text also approximates df as Δf in multiple places.

[The End of Optional Reading 2.3.1]

2.4 Tensor Fields

Definition 1 A tensor of type (k, l) on a vector space V is a multilinear map

$$T : \underbrace{V^* \times \cdots \times V^*}_{k \text{ terms}} \times \underbrace{V \times \cdots \times V}_{l \text{ terms}} \rightarrow \mathbb{R}.$$

Remark 1 T can be likened to a machine with k “upper slots” and l “lower slots”. So long as we input k dual vectors and l vectors into the upper and lower slots, respectively, this machine produces a real number which is linearly dependent on each of the inputs (this is the meaning of a “multilinear map”).

Example 1 (1) A dual vector on V is a tensor of type $(0, 1)$ on V . (2) An element of V can be regarded as a tensor of type $(1, 0)$ on V . (This is because v can be identified as v^{**} , and v^{**} is a linear map from V^* to \mathbb{R} .)

From now on, we will use $\mathcal{T}_V(k, l)$ to represent the collection of all tensors of type (k, l) on V ; thus, $V = \mathcal{T}_V(1, 0)$, $V^* = \mathcal{T}_V(0, 1)$.

Suppose $T \in \mathcal{T}_V(1, 1)$, then $T : V^* \times V \rightarrow \mathbb{R}$. However, T can also be viewed as another type of map. Since $\forall \omega \in V^*$, $v \in V$, we have $T(\omega; v) \in \mathbb{R}$, so $T(\omega; \bullet)$ is

a machine with only a lower slot that can turn a vector linearly into a real number, which means that $T(\omega; \bullet)$ is a dual vector on V , i.e., $T(\omega; \bullet) \in V^*$. After T is given, we can create $T(\omega; \bullet)$ with one $\omega \in V^*$, hence, T can also be viewed as a map (and it is linear) that turns a dual vector ω into a dual vector $T(\omega; \bullet)$, i.e., $T : V^* \xrightarrow{\text{linearly}} V^*$. Similarly, we can also view T as $T : V \xrightarrow{\text{linearly}} V$. These three viewpoints for the same $T \in \mathcal{T}_V(1, 1)$ are equivalent. For expositional convenience, we call this way of viewing the same tensor as different maps “the multifaceted view of tensors”. Being able to have a “multifaceted view” is one of the advantages of defining tensors as maps. We will use this frequently in the future.

Definition 2 The **tensor product** $T \otimes T'$ of a tensor T of type (k, l) and a tensor T' of type (k', l') on V is a tensor of type $(k + k', l + l')$ defined as follows:

$$\begin{aligned} T \otimes T'(\omega^1, \dots, \omega^k, \omega^{k+1}, \dots, \omega^{k+k'}; v_1, \dots, v_l, v_{l+1}, \dots, v_{l+l'}) \\ := T(\omega^1, \dots, \omega^k; v_1, \dots, v_l)T'(\omega^{k+1}, \dots, \omega^{k+k'}; v_{l+1}, \dots, v_{l+l'}). \end{aligned}$$

In Euclidean vector field theory, a dyadic $\vec{v}\vec{u}$ is actually the tensor product of two vectors \vec{v} and \vec{u} simply with the symbol \otimes being omitted.⁷

Do tensor products satisfy the commutative law? Suppose $\omega \in V^*$, $v \in V \equiv V^{**}$, then $v \otimes \omega \in \mathcal{T}_V(1, 1)$, $\omega \otimes v \in \mathcal{T}_V(1, 1)$. It follows from Definition 2 that $\forall \mu \in V^*$ and $u \in V$ we have $v \otimes \omega(\mu; u) = v(\mu)\omega(u) = \omega(u)v(\mu) = \omega \otimes v(\mu; u)$ [where $v(\mu)$ should be interpreted as $v^{**}(\mu)$], and hence $v \otimes \omega = \omega \otimes v$. However, the tensor product of two vectors (or two dual vectors) usually becomes another tensor after exchanging the order, i.e., $v \otimes u \neq u \otimes v$, $\omega \otimes \mu \neq \mu \otimes \omega$. For instance, a dyadic in Euclidean space does not satisfy the commutative law.

Theorem 2.4.1 $\mathcal{T}_V(k, l)$ is a vector space, with $\dim \mathcal{T}_V(k, l) = n^{k+l}$.

Proof (A) We define the addition, scalar multiplication and zero element in a natural way and make $\mathcal{T}_V(k, l)$ a vector space (see the first part of the proof of Theorem 2.3.1).

(B) Show that there are n^{k+l} basis vectors. Take $n = 2, k = 2, l = 1$ as an example (it is not difficult to prove this in the general case). Suppose $\{e_1, e_2\}$ is a basis of V , and $\{e^{1*}, e^{2*}\}$ is its dual basis. All we have to prove is that the following 8 elements form a basis of $\mathcal{T}_V(2, 1)$:

$$\begin{array}{llll} e_1 \otimes e_1 \otimes e^{1*}, & e_1 \otimes e_1 \otimes e^{2*}, & e_1 \otimes e_2 \otimes e^{1*}, & e_1 \otimes e_2 \otimes e^{2*}, \\ e_2 \otimes e_1 \otimes e^{1*}, & e_2 \otimes e_1 \otimes e^{2*}, & e_2 \otimes e_2 \otimes e^{1*}, & e_2 \otimes e_2 \otimes e^{2*}. \end{array}$$

One can first show that they are linearly independent (left as an exercise), and then show that any $T \in \mathcal{T}_V(2, 1)$ can be expressed as

⁷ Similarly, $|\psi\rangle|\phi\rangle$ in quantum mechanics is also a tensor product of $|\psi\rangle$ and $|\phi\rangle$ simply with the symbol \otimes being omitted. However, in quantum mechanics the vector space of $|\psi\rangle$ is an infinite dimensional vector space on \mathbb{C} , which is more complicated than a finite dimensional vector spaces on \mathbb{R} which we are discussing. For details see Appendix B in Volume II.

$$T = T^{\mu\nu}{}_\sigma e_\mu \otimes e_\nu \otimes e^\sigma{}^*, \quad (2.4.1)$$

where

$$T^{\mu\nu}{}_\sigma = T(e^{\mu*}, e^{\nu*}; e_\sigma). \quad (2.4.2)$$

The proof is left as an exercise. [NB: The equation to be proved, i.e., (2.4.1), is a tensor equation of type (2, 1).] \square

Remark 2 $T^{\mu\nu}{}_\sigma$ are the components of T in the basis $\{e_\mu \otimes e_\nu \otimes e^\sigma{}^*\}$, or simply “the components of T in the basis $\{e_\mu\}$ ” for short.

Now we introduce another important operation on tensors: contraction. As we just claimed, a tensor T of type (1, 1) can be regarded as a linear map from V to V , which in fact is the linear transformation we know from linear algebra. The matrix $(T^\mu{}_\nu)$ constituted by the components of T in an arbitrary basis $\{e_\mu \otimes e^\nu\}$ clearly depends on the choice of a basis, and it is not difficult to show that the two matrices $(T^\mu{}_\nu)$ and $(T'^\mu{}_\nu)$ that correspond to the components of the same T in two different bases are similar to each other; the proof is as follows. As in (2.4.2) we have

$$\begin{aligned} T'^\mu{}_\nu &= T(e'^{\mu*}; e'_\nu) = T((\tilde{A}^{-1})_\rho{}^\mu e^\rho{}^*; A^\sigma{}_\nu e_\sigma) = (\tilde{A}^{-1})_\rho{}^\mu A^\sigma{}_\nu T(e^\rho{}^*; e_\sigma) \\ &= (\tilde{A}^{-1})_\rho{}^\mu A^\sigma{}_\nu T^\rho{}_\sigma = (A^{-1})^\mu{}_\rho T^\rho{}_\sigma A^\sigma{}_\nu = (A^{-1}TA)^\mu{}_\nu, \end{aligned} \quad (2.4.3)$$

where in the first and forth step we used (2.4.2), in the second step we used Theorem 2.3.2, and in the third step we used the linearity of T . As a result, we have the matrix equation $T' = A^{-1}TA$ (where T' , A and T all represent matrices. T sometimes represents a tensor and sometimes represents a matrix, the reader should interpret it by the context). Thus we can see that T and T' are two similar matrices. Using $T'^\mu{}_\mu$ (short for $\sum_{\mu=1}^n T'^\mu{}_\mu$) and $T^\rho{}_\rho$ to represent the trace of T' and T , then from (2.3.4) we get

$$T'^\mu{}_\mu = (A^{-1})^\mu{}_\rho T^\rho{}_\sigma A^\sigma{}_\mu = A^\sigma{}_\mu (A^{-1})^\mu{}_\rho T^\rho{}_\sigma = \delta^\sigma{}_\mu T^\rho{}_\sigma = T^\rho{}_\rho.$$

This shows that a tensor of type (1, 1) has the same trace in different bases. When we are considering tensors we should pay attention to the features that do not depend on the basis, and the trace of a tensor of type (1, 1) is exactly one of these features, which is usually called the **contraction** of T , denoted by CT for now; namely,

$$CT := T^\mu{}_\mu = T(e^{\mu*}; e_\mu). \quad (2.4.4)$$

And now we discuss the contraction of a tensor T of type (2, 1). T can be denoted by $T(\bullet, \bullet; \bullet)$; it has two upper slots and one lower slot, and thus there are two possible contractions: ① The contraction on the first upper slot and the lower slot $C_1^1 T := T(e^{\mu*}, \bullet; e_\mu)$; ② The contraction on the second upper slot and the lower slot $C_1^2 T := T(\bullet, e^{\mu*}; e_\mu)$. If we define these two contractions using another basis $\{e'_\rho\}$ and denote them by $(C_1^1 T)'$ and $(C_1^2 T)'$, respectively, then it is easy to show

that (Exercise 2.14) $(C_1^1 T)' = C_1^1 T$, $(C_1^2 T)' = C_1^2 T$. From the “multifaceted view of tensors” we can see that both $C_1^1 T$ and $C_1^2 T$ are tensors of type $(1, 0)$, whose components in any basis can be expressed in terms of the components of T in this basis as $(C_1^1 T)^v = T(e^{\mu*}, e^{\nu*}; e_\mu) = T^{\mu\nu}{}_\mu$ and $(C_1^2 T)^v = T^{\nu\mu}{}_\mu$ (the summation symbol has been omitted). It is not difficult to generalize the discussion above and give a definition for the contraction of a tensor of type (k, l) as follows:

Definition 3 The **contraction** on the i th upper index ($i \leq k$) and the j th lower index ($j \leq l$) of $T \in \mathcal{T}_V(k, l)$ is defined as

$$C_j^i T := T(\bullet, \dots, e^{\mu*}, \bullet, \dots; \bullet, \dots, e_\mu, \bullet, \dots) \in \mathcal{T}_V(k-1, l-1) \quad (\text{sum over } \mu). \\ \begin{array}{ccccc} \uparrow & & \uparrow & & \\ i\text{th upper slot} & & j\text{th lower slot} & & \end{array} \quad (2.4.5)$$

Remark 3 ① $C_j^i T$ does not depend on the choice of a basis. ② It can be easily seen from (2.4.5) that any contraction of a tensor of type (k, l) is a tensor of type $(k-1, l-1)$. ③ One can construct all kinds of new tensors using tensor products in conjunction with contractions. For example, suppose $v \in V$, $\omega \in V^*$, then $v \otimes \omega$ is a tensor of type $(1, 1)$, while $C(v \otimes \omega)$ is a tensor of type $(0, 0)$ (a scalar).

Later, we will encounter the operation of contracting after taking the tensor product occurs frequently, whose conclusion can be considered as the action of a tensor on a vector (or a dual vector). As examples, here we write out three equations and then prove them.

$$(a) \quad C(v \otimes \omega) = \omega_\mu v^\mu = \omega(v) = v(\omega), \quad \forall v \in V, \omega \in V^*, \quad (2.4.6)$$

(where v^μ and ω_μ are the components of v and ω in the same basis.)

$$(b) \quad C_2^1(T \otimes v) = T(\bullet, v), \quad \forall v \in V, T \in \mathcal{T}_V(0, 2). \quad (2.4.7)$$

$$(c) \quad C_2^2(T \otimes \omega) = T(\bullet, \omega; \bullet), \quad \forall \omega \in V^*, T \in \mathcal{T}_V(2, 1). \quad (2.4.8)$$

We will only give the proof of (2.4.7), and the other two equation are left as exercises. $T \otimes v$ on the left-hand side of (2.4.7) is a tensor of type $(1, 2)$, which is a machine with 1 upper slot and 2 lower slots, and can be expressed as $T \otimes v(\bullet; \bullet, \bullet)$; hence,

$$C_2^1(T \otimes v) = T \otimes v(e^{\mu*}; \bullet, e_\mu).$$

Therefore, to prove (2.4.7) we only have to show the following equation:

$$T \otimes v(e^{\mu*}; \bullet, e_\mu) = T(\bullet, v). \quad (2.4.7')$$

Seeing that this is an equality of dual vectors, we only have to show that both sides give the same real number when applied to any $u \in V$:

$$\begin{aligned} \text{l.h.s. acting on } u &= T \otimes v(e^{\mu*}; u, e_\mu) = T(u, e_\mu)v(e^{\mu*}) \\ &= T(u, e_\mu)e^{\mu*}(v) = T(u, e_\mu)v^\mu = T(u, v) = \text{r.h.s. acting on } u, \end{aligned}$$

(where we used the result of Exercise 2.11 in the fourth equality), and thus we have (2.4.7).

Apart from the three equalities above, there are many similar ones. Those equalities represent the following rule: “The action of T on ω (or v) is contracting after taking the tensor product of T and ω (or v)”, or roughly speaking, “the action is contracting after taking product”. The manipulation of contracting after taking the tensor product of two tensors is also usually called contraction for short, and thus the expression above can even be simplified as “action means contraction”.

Now we return to a manifold M . The collection of all tensors of type (k, l) on the tangent space V_p of an arbitrary point p in M is denoted by $\mathcal{T}_{V_p}(k, l)$. Suppose $\{e_\mu\}$ and $\{e^{v*}\}$ are an arbitrary basis of V_p and its dual basis, respectively, then $T \in \mathcal{T}_{V_p}(2, 1)$ can also be written in an expanded form similar to (2.4.1). If we choose a coordinate system such that the coordinate patch contains p , then we can choose the coordinate basis vectors $\partial/\partial x^\mu$ and dual basis vectors dx^μ to be e_μ and $e^{\mu*}$; namely, we rewrite (2.4.1) as

$$T = T^{\mu\nu}_\sigma \frac{\partial}{\partial x^\mu} \otimes \frac{\partial}{\partial x^\nu} \otimes dx^\sigma, \quad (2.4.1')$$

where the coordinate components $T^{\mu\nu}_\sigma$ can be expressed following (2.4.2) as

$$T^{\mu\nu}_\sigma = T(dx^\mu, dx^\nu; \partial/\partial x^\sigma). \quad (2.4.2')$$

If we assign a tensor of type (k, l) at every point on a manifold M , we then obtain a **tensor field** of type (k, l) . A tensor field T on M is said to be **smooth** if $T(\omega^1, \dots, \omega^k; v_1, \dots, v_l) \in \mathcal{T}_M$ \forall smooth dual vector fields $\omega^1, \dots, \omega^k$ and smooth vector fields v_1, \dots, v_l . From now on, the term “tensor field” will refer to a smooth (C^∞) tensor field unless stated otherwise.

Theorem 2.4.2 *The transformation relation for the components of a tensor of type (k, l) in two coordinate systems is as follows (called the **tensor transformation law**):*

$$T'^{\mu_1 \dots \mu_k}_{\nu_1 \dots \nu_l} = \frac{\partial x'^{\mu_1}}{\partial x^{\rho_1}} \dots \frac{\partial x'^{\mu_k}}{\partial x^{\rho_k}} \frac{\partial x^{\sigma_1}}{\partial x'^{\nu_1}} \dots \frac{\partial x^{\sigma_l}}{\partial x'^{\nu_l}} T^{\rho_1 \dots \rho_k}_{\sigma_1 \dots \sigma_l}.$$

Proof Exercise. □

Remark 4 Many textbooks adopt the above equation as the definition of a tensor.

2.5 Metric Tensor Fields

Definition 1 A metric g on a vector space V is a symmetric, non-degenerate tensor of type $(0, 2)$ on V . “Symmetric” means $g(v, u) = g(u, v) \forall v, u \in V$, and “non-degenerate” means $g(v, u) = 0 \forall u \in V \Rightarrow v = 0 \in V$.

Remark 1 This abstract definition of non-degeneracy is closely related to the non-degeneracy of a matrix (that is, the determinant is nonzero) which is familiar to the reader. It can be proved that [see the paragraph after (2.6.8)] if g is non-degenerate, then the matrix constituted by the components $g_{\mu\nu} \equiv g(e_\mu, e_\nu)$ in an arbitrary basis $\{e_\mu\}$ of V is also non-degenerate. Conversely, if V has a basis in which the component matrix of g is non-degenerate, then g is non-degenerate.

A metric is similar to the familiar inner product. However, the difference between the metric above and a normal inner product is that $g(v, v)$ can be negative, and $g(v, v) = 0$ does not mean $v = 0$. Later on, we will also often call $g(v, u)$ the inner product of v and u with respect to a metric g . Once a metric is defined for a vector space V , the lengths of and the orthogonality between its elements can be defined as follows:

Definition 2 The **length** or **magnitude** of $v \in V$ is defined as $|v| := \sqrt{|g(v, v)|}$. Two vectors $v, u \in V$ are said to be **orthogonal** if $g(v, u) = 0$. A basis $\{e_\mu\}$ of V is said to be **orthonormal** if any two basis vectors are orthogonal and each basis vector e_μ satisfies $g(e_\mu, e_\mu) = \pm 1$ (not summed over μ).

Remark 2 Definition 2 indicates that the components of a metric g in any orthonormal basis satisfy

$$g_{\mu\nu} = \begin{cases} 0, & \mu \neq \nu \\ \pm 1, & \mu = \nu \end{cases}. \quad (2.5.1)$$

Thus, the matrix constituted by the components of a metric in an orthonormal basis is a diagonal matrix, and the diagonal elements are either $+1$ or -1 .

Theorem 2.5.1 Any vector space assigned with a metric has an orthonormal basis. When written as a diagonal matrix, the numbers of $+1$ and -1 among the diagonal elements do not depend on the choice of an orthonormal basis.

Proof Omitted. [See, for example, Schutz (1980) pp. 65–66]. □

Definition 3 Beyond having a diagonal matrix in an orthonormal basis, metrics whose diagonal elements are all $+1$ are said to be **positive definite** or **Riemannian**, metrics whose diagonal elements are all -1 are said to be **negative definite**, and the others are said to be **indefinite**. The indefinite metrics whose diagonal elements have only one -1 are said to be **Lorentzian**. The summation of all the diagonal elements is called the **signature** of a metric. The ones most used in relativity are the Lorentzian metrics and positive definite metrics.

Remark 3 For Lorentzian metrics, there are two conventions in the literature. Definition 3 presents the first convention, in which the diagonal elements of a 4-dimensional Lorentzian metric are $(-1, 1, 1, 1)$ (up to a trivial reordering,⁸) and the signature is $+2$. In the other convention a Lorentzian metric is defined as a metric whose diagonal elements has only one $+1$, and thus the diagonal elements of a 4-dimensional Lorentzian metric reads $(1, -1, -1, -1)$, and the signature is -2 . This text adopts the convention with the $+2$ signature.

Definition 4 There are three types of vectors in a vector space V with a Lorentzian metric g : ① any v that satisfies $g(v, v) > 0$ is called a **spacelike vector**; ② any v that satisfies $g(v, v) < 0$ is called a **timelike vector**; ③ any v that satisfies $g(v, v) = 0$ is called a **lightlike vector** or a **null vector**.

Remark 4 ① In the convention with the -2 signature, the definitions of spacelike vectors and timelike vectors are the exact opposite: a spacelike vector is defined as $g(v, v) < 0$, while a timelike vector is defined as $g(v, v) > 0$. Nonetheless, there is no essential difference: a vector that is timelike in the -2 signature is also timelike in the $+2$ signature, and vice versa. ② The zero vector is certainly a null vector, but not vice versa. Many readers may only be familiar with positive definite metrics, and may think $v = \underline{0}$ (the zero element) whenever $g(v, v) = 0$. However, if a metric is Lorentzian, then $g(v, v) = 0$ does not necessarily lead to $v = \underline{0}$ (the zero element is unique, while there are infinitely many null vectors). Nonzero 4-dimensional null vectors play a significant role in relativity. For instance, it is convenient to use them to describe the propagation of electromagnetic waves and gravitational waves.

A metric g is a tensor of type $(0, 2)$, which is a bilinear map from $V \times V$ to \mathbb{R} , so $\forall v, u \in V$ we have $g(v, u) \in \mathbb{R}$, and thus $g(v, \bullet) \in V^*$. Given g , we can create $g(v, \bullet) \in V^*$ for any $v \in V$, and hence g can be viewed as a linear map from V to V^* , i.e., $g : V \xrightarrow{\text{linearly}} V^*$, which is an isomorphism (the proof is left as Exercise 2.15). Therefore, V acquires a natural, distinguishing isomorphism from V to V^* after a metric is assigned to it, using which we can naturally identify V and V^* . Summary: V is naturally identified with V^{**} whether or not there is a metric; if there is a metric, then V can also be identified with V^* .

Now we return to a manifold M .

Definition 5 A symmetric, everywhere non-degenerate tensor field of type $(0, 2)$ is called a **metric tensor field**.

Remark 5 In this text, we only care about metric fields each of which has a signature that is the same everywhere.

One of the uses of a metric field is to define the arc length of a curve. First, we discuss a 2-dimensional Euclidean space. Suppose the parametric equation of a curve $C(t)$ in the natural coordinate system $\{x, y\}$ is $x = x(t)$, $y = y(t)$, then the square of the length of a curve segment dl^2 [short for $(dl)^2$] is

⁸ The modern convention in physics is to take “time” as the first component.

$$dl^2 = dx^2 + dy^2 = [(dx/dt)^2 + (dy/dt)^2]dt^2 = [(T^1)^2 + (T^2)^2]dt^2 = |T|^2 dt^2,$$

where T is a tangent vector of $C(T)$. From this we get

$$dl = |T|dt, \quad (2.5.2)$$

and thus we define the **arc length** of $C(t)$ as

$$l = \int |T|dt. \quad (2.5.3)$$

The equation above can be generalized to any manifold M with a positive definite metric field g . Suppose $C(t)$ is an arbitrary C^1 curve on M and T is its tangent vector, i.e., $T \equiv \partial/\partial t$, then $|T| = \sqrt{g(T, T)}$, and hence the arc length of $C(t)$ can be naturally written as

$$l := \int \sqrt{g(T, T)}dt. \quad (2.5.4)$$

For a manifold M with a Lorentzian metric field g one should pay attention to the type of a curve before defining its arc length. If the tangent vector at each point of a C^1 curve $C(t)$ is spacelike, then $C(t)$ is called a **spacelike curve**. Similarly, we can define a **timelike curve** and a **null curve**. The arc length of spacelike and null curves are also defined by (2.5.4) (and thus the arc length of a null curve is always zero). Note that for a timelike curve we have $g(T, T) < 0$, so the length of a segment of the curve is defined as $dl := \sqrt{-g(T, T)}dt$. Thus, we have the following definition:

Definition 6 Suppose a manifold M has a Lorentzian metric field g , then the arc length of a spacelike, null or timelike curve $C(t)$ can be defined as

$$l := \int \sqrt{|g(T, T)|}dt, \quad \text{where } T \equiv \frac{\partial}{\partial t}. \quad (2.5.5)$$

As for the arc length of an outlandish curve that can turn from spacelike into timelike (or the other way round), we will leave it undefined. Although the following discussion about arc length is for the Lorentzian metrics, it also applies to positive definite metrics (if we consider all curves as spacelike curves).

It is not difficult to show that (Exercise 2.16) the arc length of a curve is independent of its parametrization; that is, the reparametrization (which keeps the image unchanged and adjusts the parameter) of a curve does not change the arc length of the curve. In addition, since the definition of arc length (Definition 6) does not involve a coordinate system, the arc length is certainly independent of the coordinate system. However, if the curve lies inside the coordinate patch of a coordinate system $\{x^\mu\}$, the arc length can also be calculated with the help of the coordinate system. Since

$$g(T, T) = g(T^\mu \partial/\partial x^\mu, T^\nu \partial/\partial x^\nu) = T^\mu T^\nu g(\partial/\partial x^\mu, \partial/\partial x^\nu) = (dx^\mu/dt)(dx^\nu/dt)g_{\mu\nu},$$

[In the last step, we used that fact that “the coordinate components of a tangent vector of a curve are equal to the derivative of the parametric equation of the curve in this system with respect to the parameter” (Theorem 2.2.4), i.e., $T^\mu = dx^\mu/dt$.] the length of a line segment is

$$dl = \sqrt{|g_{\mu\nu}dx^\mu dx^\nu|}. \quad (2.5.6)$$

Introduce the notation

$$ds^2 \equiv g_{\mu\nu}dx^\mu dx^\nu, \quad (2.5.7)$$

then the arc length reads

$$l = \int \sqrt{ds^2} \quad (\text{for spacelike curves}), \quad (2.5.8)$$

$$l = \int \sqrt{-ds^2} \quad (\text{for timelike curves}). \quad (2.5.9)$$

The notation ds^2 shows up very frequently in differential geometry; it is usually called a **line element**. For a spacelike curve, ds^2 is equal to dl^2 , i.e., the square of the length of a line segment dl ; for a timelike curve, ds^2 is equal to $-dl^2$, and thus it is not the square of any real number. In fact, ds^2 is just a notation defined by (2.5.7), which is not the square of any real number at all for any timelike curve [see Optional Reading 2.5.2 for a precise interpretation of (2.5.7)]. However, since the right-hand side of $ds^2 \equiv g_{\mu\nu}dx^\mu dx^\nu$ contains all the components $g_{\mu\nu}$ of g in the coordinate system involved, one can “read off” all the coordinate components of the metric directly from the expression of the line element. For example, suppose the expression for the line element of a metric g on a 2-dimensional manifold in a coordinate system $\{x, t\}$ is

$$ds^2 = -xdt^2 + dx^2 + 4dtdx, \quad (2.5.10)$$

then we can read off the components of g in this system as $g_{tt} = -x$, $g_{xx} = 1$, $g_{tx} = g_{xt} = 2$. Thus, we can see that a given line element (expression) is equivalent to the given metric field.

Suppose $C : I \rightarrow M$ is a spacelike or timelike curve, then $|T|$, the length of the tangent vector T at an arbitrary point $C(t)$, is a function of t denoted by $|T|(t)$. If we assign a point $C(t_0)$ on the curve arbitrarily as the starting point for measuring length, then the curve segment between $C(t_0)$ and $C(t)$ has the length $l(t) = \int_{t_0}^t |T|(t')dt'$, which is a increasing function of t . Hence, l can also act as the parameter of this curve, called the **arc length parameter**. From $dl \equiv \sqrt{|g(T, T)|}dt$ we can see that a tangent vector of a curve with the arc length as its parameter satisfies $|g(T, T)| = 1$, namely it has a unit length.

Definition 7 Suppose a metric field g is given on a manifold M , then (M, g) is called a **generalized Riemannian space**. (If g is positive definite, it is called a **Riemannian**

space; if g is Lorentzian, it is called a **pseudo-Riemannian space**, or in physics, it is called a **spacetime**.⁹⁾

Now, we introduce two simple but significant examples of generalized Riemannian spaces, namely Euclidean space and Minkowski space.

Definition 8 Suppose x^μ are the natural coordinates of \mathbb{R}^n . Define a metric tensor field δ on \mathbb{R}^n as

$$\delta := \delta_{\mu\nu} dx^\mu \otimes dx^\nu, \quad (2.5.11)$$

then (\mathbb{R}^n, δ) is called the **n -dimensional Euclidean space**, and δ is called the **Euclidean metric**.

The equation above indicates that the components of δ in a dual coordinate basis of the natural coordinate system are

$$\delta_{\mu\nu} = \begin{cases} 0, & \mu \neq \nu \\ +1, & \mu = \nu \end{cases}.$$

Therefore, according to (2.5.7), the expression for the line element of the Euclidean metric in the natural coordinate system should be $ds^2 = \delta_{\mu\nu} dx^\mu dx^\nu$. If $n = 2$, then we have $ds^2 = (dx^1)^2 + (dx^2)^2$. This is exactly the well-known expression for the line element of the 2-dimensional Euclidean space. It follows from (2.5.11) that the natural coordinate basis is orthonormal measured by the Euclidean metric, since from

$$\delta(\partial/\partial x^\alpha, \partial/\partial x^\beta) = \delta_{\mu\nu} dx^\mu \otimes dx^\nu (\partial/\partial x^\alpha, \partial/\partial x^\beta) = \delta_{\mu\nu} dx^\mu (\partial/\partial x^\alpha) dx^\nu (\partial/\partial x^\beta)$$

we can easily see that

$$\delta(\partial/\partial x^\alpha, \partial/\partial x^\beta) = \delta_{\alpha\beta}. \quad (2.5.12)$$

However, a coordinate system that satisfies (2.5.12) is not necessarily the natural coordinate system. For example, for 2-dimensional Euclidean space, the coordinate system defined based on the natural coordinate system $\{x, y\}$ as follows

$$x' = x + a, \quad y' = y + b \quad (a, b \text{ are constants}) \quad (2.5.13)$$

has a basis $\{\partial/\partial x', \partial/\partial y'\}$ that satisfies (2.5.12) (and thus it is orthonormal). Furthermore, it is not difficult to show that (Exercise 2.17) the coordinate bases $\{\partial/\partial x', \partial/\partial y'\}$ of $\{x', y'\}$ defined by the following three equations also satisfy (2.5.12):

⁹ More precisely, (M, g) is called a **spacetime** if M is a *connected* manifold, and g is a Lorentzian metric field with adequate differentiability.

$$x' = x \cos \alpha + y \sin \alpha, \quad y' = -x \sin \alpha + y \cos \alpha \quad (\alpha \text{ is a constant}), \quad (2.5.14)$$

$$x' = -x, \quad y' = y, \quad (2.5.15)$$

$$x' = x, \quad y' = -y. \quad (2.5.16)$$

Definition 9 A coordinate system in an n -dimensional Euclidean space that satisfies (2.5.12) is called a **Cartesian coordinate system** or **rectangular coordinate system**. In other words, a coordinate system is called a Cartesian system if its coordinate basis is orthonormal measured by the Euclidean metric δ .

Remark 6 ① Since (2.5.12) is equivalent to (2.5.11), one can also say that a coordinate system that satisfies (2.5.11) is a Cartesian system. ② The natural coordinate system is certainly a Cartesian system. ③ The relationship between any two Cartesian systems in 2-dimensional Euclidean space can only have one of the forms in (2.5.13)–(2.5.16) (or a composite of them). The first is called a **translation**, the second is called a **rotation**, and each of the final two is called a **reflection**. ④ One should distinguish between the symbols δ and $\delta_{\mu\nu}$. δ stands for the Euclidean metric field, which is a tensor field, while $\delta_{\mu\nu}$ are the components of δ in a Cartesian system. Also note that the components of δ in a non-Cartesian system are not $\delta_{\mu\nu}$.

The polar coordinate system $\{r, \varphi\}$ is an example of a non-Cartesian system in 2-dimensional Euclidean space. When using the polar coordinate system, physics books usually use $\{\hat{e}_r, \hat{e}_\varphi\}$ as the corresponding basis (the \wedge above stands for unit vectors), which is an orthonormal basis. However, it is not the coordinate basis of the polar coordinate system; the point is that $\partial/\partial\varphi$ is not normalized since $\delta(\partial/\partial\varphi, \partial/\partial\varphi) = r^2 \neq 1$ (the proof is left as an exercise). In fact, \hat{e}_φ is the result of normalizing $\partial/\partial\varphi$, i.e., $\hat{e}_\varphi := r^{-1}\partial/\partial\varphi$. Thus, the frequently used basis $\{\hat{e}_r, \hat{e}_\varphi\}$ in physics books is not the coordinate basis of the polar coordinate system but rather it is the orthonormal basis that corresponds to the polar coordinate system.

Euclidean space is the simplest Riemannian space. Now we introduce the simplest pseudo-Riemannian space—Minkowski space. The diagonal elements of the diagonalized 4-dimensional Lorentzian metric are $(-1, 1, 1, 1)$. To highlight the only -1 , we denote its position as row 0 and column 0, and denote the position of the three $+1$ s as rows 1, 2, 3 and columns 1, 2, 3. Denote the elements of this diagonal matrix as $\eta_{\mu\nu}$ (to distinguish it from $\delta_{\mu\nu}$), i.e., $\eta_{00} \equiv -1$, $\eta_{11} \equiv \eta_{22} \equiv \eta_{33} \equiv 1$. Generalizing it to n dimensions, we have

$$\eta_{\mu\nu} = \begin{cases} 0, & \mu \neq \nu \\ -1, & \mu = \nu = 0, \\ +1, & \mu = \nu = 1, \dots, n-1. \end{cases}$$

Now we present the definition of Minkowski space.

Definition 10 Suppose x^μ are the natural coordinates of \mathbb{R}^n . Define a metric tensor field η on \mathbb{R}^n as

$$\eta := \eta_{\mu\nu} dx^\mu \otimes dx^\nu , \quad (2.5.17)$$

then (\mathbb{R}^n, η) is called the **n -dimensional Minkowski space** (also known as the **n -dimensional Minkowski spacetime** in physics), and η is called the **Minkowski metric**.

From Definition 10 we can see that the expression for the line element of Minkowski space in the natural coordinate system is $ds^2 = \eta_{\mu\nu} dx^\mu dx^\nu$. Take $n = 4$, for example, we have $ds^2 = -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2$. This is exactly the well-known expression for the line element of 4-dimensional Minkowski spacetime. It is easy to show that

$$\eta(\partial/\partial x^\alpha, \partial/\partial x^\beta) = \eta_{\alpha\beta} , \quad (2.5.18)$$

and thus the natural coordinate basis $\{\partial/\partial x^\mu\}$ is also orthonormal as measured by the Minkowski metric. (The 0th coordinate basis vector is normalized to -1 , the others are normalized to 1). However, a basis satisfying (2.5.18) is not necessarily the basis of the natural coordinate system. For instance, suppose t and x are the natural coordinates for 2-dimensional Minkowski space, then the coordinate basis $\{\partial/\partial t', \partial/\partial x'\}$ of

$$t' = t + a , \quad x' = x + b \quad (a, b \text{ are constants}) \quad (2.5.19)$$

also satisfies (2.5.18). It is not difficult to verify that (Exercise 2.18) the coordinate bases $\{\partial/\partial t', \partial/\partial x'\}$ of $\{t', x'\}$ defined by the following three equations also satisfy (2.5.18):

$$t' = t \cosh \lambda + y \sinh \lambda , \quad x' = t \sinh \lambda + x \cosh \lambda \quad (\alpha \text{ is a constant}) , \quad (2.5.20)$$

$$t' = -t , \quad x' = x , \quad (2.5.21)$$

$$t' = t , \quad x' = -x . \quad (2.5.22)$$

Definition 11 The coordinate system in the n -dimensional Minkowski space that satisfies (2.5.18) is called a **Lorentzian coordinate system** or **pseudo-Cartesian coordinate system**; some works may also refer to it as a Cartesian (or Minkowski) coordinate system.

Remark 7 ① The natural coordinates of Minkowski space are certainly Lorentzian coordinates. ② The relationship between any two Lorentzian coordinate systems in 2-dimensional Minkowski space can only have one of the forms as in (2.5.19)–(2.5.22) (or a composite of them). The first one is called a **translation**, the second one (2.5.20) is called a **boost**, and each of the final two is called a **reflection**. ③ The components of the Minkowski metric tensor η in a non-Lorentzian coordinate basis are not equal to $\eta_{\mu\nu}$.

[Optional Reading 2.5.1]

Unlike any translation, rotation or boost, a reflection is a “discrete” transformation. There is another kind of “discrete” transformation called **inversion**, defined as $x' = -x$, $y' = -y$ and $t' = -t$, $x' = -x$ for 2-dimensional Euclidean space and Minkowski space, respectively. Unlike a reflection, an inversion is a symmetry transformation with respect to a point. However, an inversion is not an independent transformation. Specifically, $x' = -x$, $y' = -y$ is a special case of (2.5.14) when $\alpha = \pi$, while $t' = -t$, $x' = -x$ can be regarded as a composite of (2.5.21) and (2.5.22).

[The End of Optional Reading 2.5.1]

[Optional Reading 2.5.2]

In the text above, we have interpreted dl^2 as the square of the length of a line segment. This is just a “popular” interpretation used a lot by physicists, which is actually not precise. For instance, the 2-dimensional Euclidean space has

$$dl^2 = dx^2 + dy^2. \quad (2.5.23)$$

If a curve is a straight line, the essence of (2.5.23) is then

$$(\Delta l)^2 = (\Delta x)^2 + (\Delta y)^2, \quad (2.5.24)$$

where Δl is a finite segment from the straight line, and Δx and Δy are the finite increments of the coordinates x and y for this segment, respectively. If the curve is not a straight line, then (2.5.24) is not applicable. However, no matter how short the segment is (as long as the two ends are not coincident), its length is a definite real number rather than an infinitesimal, while the prerequisite of (2.5.23) is that dl is an infinitesimal (and thus so are dx and dy). Again, we run into the trouble of an “infinitesimal nonzero quantity” (see Optional Reading 2.3.1). The same problem also occurs for the dl^2 and ds^2 of a curved space. Physicists are accustomed to using approximation methods to deal with similar issues, they (including the authors of this text) may write (2.5.23), but interpret dl , dx , dy , etc. as definite nonzero quantities Δl , Δx , Δy , etc., i.e., interpret (2.5.23) as (2.5.24). Now, we will discuss how to understand the differential geometry generalization of (2.5.23), i.e., the expression for the line element

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu. \quad (2.5.25)$$

Since both dx^μ and dx^ν are both dual vectors, their “product” $dx^\mu dx^\nu$ can only be the tensor product “ $dx^\mu \otimes dx^\nu$ ”, thus, the right-hand side of (2.5.25) is actually an abbreviation for $g_{\mu\nu} dx^\mu \otimes dx^\nu$. However, $g_{\mu\nu} dx^\mu \otimes dx^\nu$ is nothing but the expansion of the metric tensor g in the dual coordinate basis, i.e.,

$$g = g_{\mu\nu} dx^\mu \otimes dx^\nu. \quad (2.5.26)$$

On the other hand, in differential geometry, one cannot find any other interpretation for ds^2 on the left-hand side of (2.5.25), it is actually nothing but another notation for g ! Therefore, we can see that the precise meaning of (2.5.25) turns out to be the tensor equation (2.5.26). This interpretation is accurate, but also sounds pedantic, and is hard to be popularized. In contrast, one of the important reasons why (2.5.25) is commonly used is that when using approximations, dl^2 can be viewed as the square of the length of a line segment, and ds^2 is nothing but a notation for dl^2 (for spacelike segments) or $-dl^2$ (for timelike segments). Many equations in this section can only be understood with this interpretation of approximation. For instance, if we insist that we use the true, precise definition from differential geometry, then (2.5.8) should be rewritten as

$$l = \int \sqrt{g_{\mu\nu} \frac{dx^\mu}{dt} \frac{dx^\nu}{dt}} dt, \quad (2.5.8')$$

where t is the parameter of the curve we are talking about. Unlike (2.5.8), each symbol in this equation has a precise meaning; for example, dx^μ/dt is the μ th coordinate component of a tangent vector of the curve, while dt together with the integral sign indicate that the variable of integration is t .

[The End of Optional Reading 2.5.2]

2.6 The Abstract Index Notation

There are two common ways to express a tensor. The first one is using a letter without any index (such as T) to represent a tensor, though this contains two drawbacks: ① one cannot tell the type of a tensor; ② it is not easy to state that a contraction is between which upper slot and which lower slot. (The symbol $C_j^i T$ we used before is only temporary, it is not convenient to use in computations.) The second notation is to use the components (such as $T^{\mu\nu\rho}$) to represent a tensor, and to use the equalities obeyed by the components to represent the equalities obeyed by tensors. The equalities of components are the equalities of numbers, and thus all of the tensor equations in the literature using this notation are equalities of numbers. This notation can overcome the two difficulties of the first notation; however, it has a serious disadvantage of itself: sometimes, one can choose a special basis and obtain a relatively simple equation relating its components, but this equation only holds for this basis, and cannot be used to represent the tensor equation in general. We want to know which equations can and which cannot represent tensor equations, yet this is difficult to tell in this component notation. To overcome this problem, Roger Penrose created the “abstract index notation”. The main points are as follows:

1. A tensor of type (k, l) is represented by a letter with k upper indices and l lower indices, all the indices are lower-case Latin letters, which only indicate the type of a tensor, and thus are called **abstract indices**. For example, v^a stands for a vector, in which the upper index a plays the same role as the \rightarrow in \vec{v} (and hence one cannot say $a = 1$ or $a = 2$), ω_a stands for a dual vector, T^{ab}_c stands for a tensor of type $(2, 1)$, and so on. v^b and v^a stand for the same vector (i.e., \vec{v}); however, we should pay attention to the “balance of indices” when writing an equation. For example, one can write $\alpha u^a + v^a = w^a$ or $\alpha u^b + v^b = w^b$, but not $\alpha u^a + v^b = w^a$.

2. Repeated upper and lower indices represent the contraction between these two indices; for example,

$$T^a{}_a = T(e^{\mu*}; e_\mu) = T^\mu{}_\mu, \quad T^{ab}{}_a = T(e^{\mu*}, \bullet; e_\mu), \quad T^{ab}{}_b = T(\bullet, e^{\mu*}; e_\mu).$$

3. The tensor product symbol is omitted. For instance, suppose $T \in \mathcal{T}_V(2, 1)$, $S \in \mathcal{T}_V(1, 1)$, then $T \otimes S$ can be written as $T^{ab}{}_c S^d{}_e$. In the notation without indices, generally, $\omega \otimes \mu \neq \mu \otimes \omega$, as when acting on (v, u) , whether ω acts on u or v depends on the order of these letters [the first letter in $\omega \otimes \mu$ act on the first letter

in (v, u) , i.e., ω acts on v]. In the abstract index notation, since repeated upper and lower indices are assumed to be contracted, $\omega \otimes \mu(v, u)$ can be written as either $\omega_a \mu_b v^a u^b$ or $\mu_b \omega_a v^a u^b$ [both stand for $\omega(v)\mu(u)$]. Since the acting target of both $\omega_a \mu_b$ and $\mu_b \omega_a$ is the same $v^a u^b$, we have $\omega_a \mu_b = \mu_b \omega_a$. That is, the letters that represent tensors can be interchanged assuming their indices travel with them. The non-commutativity of the order of a tensor product is now manifested by $\omega_a \mu_b \neq \omega_b \mu_a$.

4. When we are talking about the components of a tensor, the corresponding indices are labeled by lower-case Greek letters, such as μ, ν, α, β , etc. (as we used before). These indices are called **component indices** or **concrete indices**, and we can ask about whether $\mu = 1$ or $\mu = 2$. A basis expansion of a tensor $T = T^{\mu\nu}{}_\sigma e_\mu \otimes e_\nu \otimes e^{\sigma*}$ can now be written as

$$T^{ab}{}_c = T^{\mu\nu}{}_\sigma (e_\mu)^a (e_\nu)^b (e^\sigma)_c, \quad (2.6.1)$$

[the lower index c of $(e^\sigma)_c$ has already indicated that it is a dual basis vector, so there is no need to write $(e^{\sigma*})_c$] while $T^{\mu\nu}{}_\sigma = T(e^{\mu*}, e^{\nu*}; e_\sigma)$ can now be written as

$$T^{\mu\nu}{}_\sigma = T^{ab}{}_c (e^\mu)_a (e^\nu)_b (e_\sigma)^c. \quad (2.6.2)$$

Note that the indices of both (2.6.1) and (2.6.2) (whether abstract or concrete) are “balanced”. Suppose $T \in \mathcal{T}_V(0, 2)$, then T should be denoted by T_{ab} . Let e_μ be the μ th basis vector of a basis, then from (2.4.7) we can see that $T(\bullet, e_\mu) = C_2^1(T \otimes e_\mu)$, and since $T \otimes e_\mu$ should be denoted by $T_{ab}(e_\mu)^c$ using the abstract index notation, $T(\bullet, e_\mu)$ should be denoted by $T_{ab}(e_\mu)^b$, also abbreviated as $T_{a\mu}$, i.e.,

$$T(\bullet, e_\mu) \equiv T_{ab}(e_\mu)^b = T_{a\mu}. \quad (2.6.3)$$

This is an expression with both abstract and component indices; we may consider T_{a1}, \dots, T_{an} as n dual vectors, where $T_{a\mu}$ stands for “the μ th dual vector”.

5. From the “multifaceted view of tensors”, we can see that a tensor of type $(1, 1)$ $T^a{}_b$ on V can be viewed either as a linear map from V to V or a linear map from V^* to V^* . That is, $T^a{}_b$ acting on a vector $v^b \in V$ still returns a vector, denoted by $u^a \equiv T^a{}_b v^b \in V$, while $T^a{}_b$ acting on a dual vector $\omega_a \in V^*$ still returns a dual vector, denoted by $\mu_b \equiv T^a{}_b \omega_a \in V^*$. Actually, it can be seen at a glance from the abstract index notation that $T^a{}_b v^b$ and $T^a{}_b \omega_a$ are a vector and a dual vector, respectively. Thus, the abstract index notation is a simple and intuitive representation of the “multifaceted view of tensors”. Using $\delta^a{}_b$ to represent the identity map from V to V , i.e., $\delta^a{}_b v^b := v^a \forall v^b \in V$, we can easily see that it is also an identity map from V^* to V^* , i.e., $\delta^a{}_b \omega_a = \omega_b \forall \omega_a \in V^*$. It is not difficult to further show that (exercise) the result of $\delta^a{}_b$ contracting with any tensor is substituting the upper index b of that tensor with a (or substituting the lower index a with b), such as $\delta^a{}_b T_{ac} = T_{bc}$, $\delta^a{}_b T^{cb}{}_e = T^{ca}{}_e$. Suppose $\{(e_\mu)^a\}$ is a basis of V , and $\{(e^\mu)_a\}$ is the dual basis, then

$$(e^\mu)_a (e_\mu)^b = \delta^b{}_a. \quad (2.6.4)$$

This is a tensor of type $(1, 1)$; to prove it, we only need to verify that the result of each side acting on an arbitrary vector v^a is the same (exercise). Suppose $\{(e_\mu)^a\}$ is a basis of V and $\{(e^\mu)_a\}$ is the dual basis, then the components of $\delta^a{}_b$ in this basis $\delta^\mu{}_v \equiv \delta^a{}_b (e^\mu)_a (e_v)^b$ satisfy $\delta^\mu{}_v = \begin{cases} +1, & (\mu = v) \\ 0, & (\mu \neq v) \end{cases}$. The proof is very simple: taking $\delta^1{}_1$ as an example, $\delta^1{}_1 = \delta^a{}_b (e^1)_a (e_1)^b = (e^1)_a (e_1)^a = 1$. Note that $\delta^0{}_0 = +1$ even for the Lorentzian signature.

6. Since a metric $g \in \mathcal{T}_V(0, 2)$, it should be denoted by g_{ab} . Suppose $v \in V$, then $g(\bullet, v) \in V^*$ (see the paragraph after Example 1 in Sect. 2.4). Regarding g as the T in (2.4.7), we get $g(\bullet, v) = C_2^1(g \otimes v) = C_2^1(g_{ab} v^c) = g_{ab} v^b$; hence, $g(\bullet, v)$ should be denoted by $g_{ab} v^b$. Further, when there is a metric g , V is identified with V^* under the isomorphism $g : V \rightarrow V^*$, and $g_{ab} v^b \equiv g(\bullet, v)$ is exactly the image of v^a under this map. Hence $g_{ab} v^b$ should be identified with v^a , and may just simply be denoted by v_a (which can be taken as a definition of v_a). That is, although mathematically speaking v^a and v_a are two different types of objects (a vector and a dual vector), in application they represent the same thing (and thus both are denoted by v). Thus, we usually write

$$v_a = g_{ab} v^b. \quad (2.6.5)$$

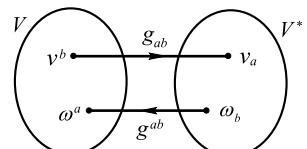
On the other hand, since $g : V \rightarrow V^*$ is an isomorphism, its inverse map g^{-1} naturally exists. It is not difficult to show that g^{-1} is a tensor of type $(2, 0)$. Though it seems it should be denoted by $(g^{-1})^{ab}$, it is usually denoted by g^{ab} (the upper indices prevent confusion with g_{ab}). Using similar reasoning, the image of any $\omega_b \in V^*$ under a map g^{ab} is $g^{ab} \omega_b$, which may simply be denoted by ω^a in order to indicate that it represents the same thing as ω_a ; therefore, (see Fig. 2.10)

$$\omega^a = g^{ab} \omega_b. \quad (2.6.6)$$

Equations (2.6.5) and (2.6.6) indicate that one can use g_{ab} and g^{ab} to “raise” and “lower” the upper and lower indices, respectively. These operations of raising and lowering indices are applicable for any abstract index in any tensor. For instance, a tensor T of type $(1, 1)$ can be denoted by $T^a{}_b$ in abstract index notation, and lowering the index using the metric is actually performing the tensor product and contraction between g and T to obtain a tensor of type $(0, 2)$, $g(\bullet, e_\mu) \otimes T(e^{\mu*}, \bullet)$, which is denoted by T_{ab} in abstract index notation, i.e., $T_{ab} \equiv g_{ac} T^c{}_b$.

Using (2.6.6) and (2.6.5) in turn we have

Fig. 2.10 A metric g can naturally identify V with V^* , and thus the indices may be raised and lowered



$$\omega^a = g^{ab}\omega_b = g^{ab}(g_{bc}\omega^c), \quad \forall \omega^a \in V,$$

and hence

$$g^{ab}g_{bc} = \delta^a{}_c, \quad (2.6.7)$$

which is actually a corollary of the fact that g^{ab} is the inverse of g_{ab} .

Suppose $\{(e_\mu)^a\}$ is a basis of V , and $\{(e^\mu)_a\}$ is the dual basis, use $g_{\mu\nu}$ and $g^{\mu\nu}$ to represent the components of g_{ab} and g^{ab} in this basis, respectively, then on the one hand we have $\delta^a{}_c = \delta^\mu{}_\sigma (e_\mu)^a (e^\sigma)_c$, on the other hand

$$\delta^a{}_c = g^{ab}g_{bc} = g^{\mu\nu}(e_\mu)^a(e_\nu)^b g_{\rho\sigma}(e^\rho)_b (e^\sigma)^c = g^{\mu\nu}g_{\nu\sigma}(e_\mu)^a(e^\sigma)_c,$$

where the third equality is because $(e_\nu)^b g_{\rho\sigma}(e^\rho)_b = \delta^\rho{}_\nu g_{\rho\sigma} = g_{\nu\sigma}$, and hence

$$g^{\mu\nu}g_{\nu\sigma} = \delta^\mu{}_\sigma. \quad (2.6.8)$$

The above equation indicates that the matrix formed by the components $g_{\mu\nu}$ of the metric g_{ab} in any basis is invertible (whose inverse is the matrix formed by the components $g^{\mu\nu}$ of the inverse metric g^{ab} in the same basis), and thus is non-degenerate. Therefore, the non-degeneracy of g_{ab} assures the non-degeneracy of its matrix $(g_{\mu\nu})$ in any basis. Conversely, suppose there exist a basis $\{(e_\mu)^a\}$ and its dual basis $\{(e^\mu)_a\}$ such that $(g_{\mu\nu})$ is non-degenerate, then $(g_{\mu\nu})$ has an inverse matrix $(g^{\mu\nu})$. Let $g^{ab} \equiv g^{\mu\nu}(e_\mu)^a(e_\nu)^b$, then it is easy to prove from $g^{\mu\nu}g_{\nu\sigma} = \delta^\mu{}_\sigma$ that $g^{ab}g_{bc} = \delta^a{}_c$, and thus $g_{ab} : V \rightarrow V^*$ is non-degenerate since it has an inverse map g^{ab} . (The proof of “the inverse exists \Rightarrow non-degenerate” is left as an exercise. Hint: $g_{ab} : V \rightarrow V^*$ having an inverse indicates that it is a one-to-one map, while if g_{ab} is degenerate, then, besides the zero element, there would be a $v^a \neq 0$ in V whose image is also $0 \in V^*$, which contradicts the fact that g_{ab} is one-to-one.)

It is not difficult to see that the upper and lower indices of the components of a tensor can be raised and lowered using the components of a metric $g_{\mu\nu}$ and its inverse $g^{\mu\nu}$. For instance, we can write $g_{\mu\nu}v^\nu$ as v_μ because

$$g_{\mu\nu}v^\nu = g_{ab}(e_\mu)^a(e_\nu)^b v^\nu = g_{ab}(e_\mu)^a v^b = v_a(e_\mu)^a = v_\mu.$$

As an example of the abstract index notation, here we introduce the abstract index expression for the 4-dimensional Minkowski metric η_{ab} .

The definition of the Minkowski metric (2.5.17) can be expressed in abstract index notation as

$$\eta_{ab} := \eta_{\mu\nu}(dx^\mu)_a(dx^\nu)_b,$$

where $\{(dx^\mu)_a\}$ is the dual basis of the Lorentzian coordinate system. If we use $\{t, x, y, z\}$ to represent $\{x^0, x^1, x^2, x^3\}$, then since the only nonzero $\eta_{\mu\nu}$ are $\eta_{00} = -1$ and $\eta_{11} = \eta_{22} = \eta_{33} = 1$, the equation above can be expressed as

$$\eta_{ab} = -(dt)_a(dt)_b + (dx)_a(dx)_b + (dy)_a(dy)_b + (dz)_a(dz)_b, \quad (2.6.9a)$$

which corresponds to the expression for the line element $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$. If we use the spherical coordinate system $\{t, r, \theta, \varphi\}$ instead, then using

$$x = r \sin \theta \cos \varphi, y = r \sin \theta \sin \varphi, z = r \cos \theta,$$

it is not difficult to derive from (2.6.9a) that

$$\eta_{ab} = -(dt)_a(dt)_b + (dr)_a(dr)_b + r^2(d\theta)_a(d\theta)_b + r^2 \sin^2 \theta (d\varphi)_a(d\varphi)_b, \quad (2.6.9b)$$

which corresponds to the line element $ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2)$.

In much of the literature that does not use the abstract index notation, the component indices in a 4-dimensional spacetime and a 3-dimensional Riemannian space are denoted by Greek letters μ, ν, \dots (each can be 0, 1, 2, 3) and Latin letters i, j, k, \dots (each can be 1, 2, 3), respectively. According to what we mentioned previously, the Latin indices in this text are supposed to represent abstract indices. However, in order to distinguish the component indices of 4 dimensions and 3 dimensions, we allow one exception: whenever we discuss a 3-dimensional Riemannian space, the Latin letters that start from i (i, j, k, \dots) are component indices (each can be 1, 2, 3), and the other Latin letters (such as a, b, c , etc.) are still abstract indices. For example, a 3-dimensional vector \vec{v} can be expressed as $v^a = v^i(\partial/\partial x^i)^a$ (i is summed from 1 to 3).

In the abstract index notation, a coordinate basis vector is denoted by $(\partial/\partial x^\mu)^a$, and a dual coordinate basis vector is denoted by $(dx^\mu)_a$. Using a metric g_{ab} and its inverse g^{ab} to raise and lower their indices, respectively, we obtain a dual vector $g_{ab}(\partial/\partial x^\mu)^b$ and a vector $g^{ab}(dx^\mu)_b$. Denote the $g_{ab}(\partial/\partial x^\mu)^b$ by ω_a for short and expand it using the dual coordinate basis as $g_{ab}(\partial/\partial x^\mu)^b = \omega_v(dx^v)_a$. Applying both sides to $(\partial/\partial x^\sigma)^a$ yields $g_{\sigma\mu} = \omega_\sigma$; hence,

$$g_{ab}(\partial/\partial x^\mu)^b = g_{\mu\nu}(dx^\nu)_a. \quad (2.6.10a)$$

Thus, in general $g^{ab}(dx^\mu)_b$ does not equal $(dx^\mu)_a$. Similarly, we have

$$g^{ab}(dx^\mu)_b = g^{\mu\nu}(\partial/\partial x^\nu)^a. \quad (2.6.10b)$$

When $g_{ab} = \delta_{ab}$ (Euclidean metric) and $\{x^\mu\}$ is a Cartesian coordinate system, these two equations above can be simplified as

$$\delta_{ab}(\partial/\partial x^\mu)^b = (dx^\mu)_a, \quad \delta^{ab}(dx^\mu)_b = (\partial/\partial x^\mu)^a, \quad (2.6.11)$$

and when $g_{ab} = \eta_{ab}$ (take 4-dimensional Minkowski as an example) and $\{x^\mu\}$ is a Lorentzian coordinate system, then we have

$$\eta_{ab}(\partial/\partial x^0)^b = -(dx^0)_a, \quad \eta_{ab}(\partial/\partial x^i)^b = (dx^i)_a; \quad (2.6.12a)$$

$$\eta^{ab}(dx^0)_b = -(\partial/\partial x^0)^a, \quad \eta^{ab}(dx^i)_b = (\partial/\partial x^i)^a. \quad (2.6.12b)$$

Here, $i = 1, 2, 3$ are not abstract indices.

An upper index and a lower index are also called a **contravariant index** and a **covariant index**, respectively. Correspondingly, a vector v^a and a dual vector ω_a are also called a **contravariant vector** and a **covariant vector**, respectively.

The symmetry of a tensor can be expressed conveniently as follows using the abstract index notation.

Definition 1 $T \in \mathcal{T}_V(0, 2)$ is said to be **symmetric** if $T(u, v) = T(v, u)$, $\forall u, v \in V$.

Since $T(u, v) = T_{ab}u^a v^b$ and $T(v, u) = T_{ab}v^a u^b = T_{ba}u^a v^b$, in abstract index notation, the necessary and sufficient condition of T to be symmetric is that $T_{ab} = T_{ba}$. In the abstract index notation, a tensor of type $(0, 2)$ can be denoted as either T_{ab} or T_{ba} . However, only when T is symmetric can we write $T_{ab} = T_{ba}$; thus, one should be more careful when writing an equation than when writing a tensor using abstract indices. Similarly, a tensor of type $(1, 1)$ can be expressed as either $T^a{}_b$ or $T_b{}^a$, whose upper indices can be lowered using a metric as $g_{ca}T^a{}_b = T_{cb}$ and $g_{ca}T_b{}^a = T_{bc}$. Although they stand for the same tensor, only a tensor of type $(1, 1)$ that is symmetric after the index is lowered can be written as $T^a{}_b = T_b{}^a$. When the indices are not raised or lowered using a metric, the upper and lower indices of a tensor of type (k, l) are ordered separately, and there is no order between an upper index and a lower index. Therefore, if we want, a tensor of type $(1, 1)$ can be written as $T_b{}^a$, a tensor of type $(2, 1)$ can be written as $T_c{}^{ab}$, etc. However, uncertainty will occur when raising and lowering indices in this kind of expression. Since we will raise and lower indices frequently, in this text we stagger the upper and lower indices from the very beginning, e.g., $T^{ab}{}_c$.

The discussion above indicates that abstract index notation is formally quite similar to the component index notation. This is exactly one of the merits of the abstract index notation: it can represent tensor equations yet retains many advantages of the component index notation.

Definition 2 For a tensor T_{ab} of type $(0, 2)$, the **symmetric part** (denoted by $T_{(ab)}$) and the **antisymmetric part** ($T_{[ab]}$) are defined respectively as

$$T_{(ab)} := \frac{1}{2}(T_{ab} + T_{ba}), \quad T_{[ab]} := \frac{1}{2}(T_{ab} - T_{ba}),$$

generally, the symmetric and antisymmetric parts of a tensor $T_{a_1 \dots a_l}$ of type $(0, l)$ are defined as

$$T_{(a_1 \dots a_l)} := \frac{1}{l!} \sum_{\pi} T_{a_{\pi(1)} \dots a_{\pi(l)}}, \tag{2.6.13}$$

$$T_{[a_1 \dots a_l]} := \frac{1}{l!} \sum_{\pi} \delta_{\pi} T_{a_{\pi(1)} \dots a_{\pi(l)}}, \tag{2.6.14}$$

where π represents a permutation of $(1, \dots, l)$, $\pi(1)$ stands for the first number in the permutation described by π , \sum_{π} represents the summation of all permutations,

and $\delta_\pi \equiv \pm 1$ (+ for even permutations, – for odd permutations). For example,

$$\begin{aligned} T_{(a_1 a_2 a_3)} &:= \frac{1}{6} (T_{a_1 a_2 a_3} + T_{a_3 a_1 a_2} + T_{a_2 a_3 a_1} + T_{a_1 a_3 a_2} + T_{a_3 a_2 a_1} + T_{a_2 a_1 a_3}), \\ T_{[a_1 a_2 a_3]} &:= \frac{1}{6} (T_{a_1 a_2 a_3} + T_{a_3 a_1 a_2} + T_{a_2 a_3 a_1} - T_{a_1 a_3 a_2} - T_{a_3 a_2 a_1} - T_{a_2 a_1 a_3}). \end{aligned}$$

Definition 3 $T \in \mathcal{T}_V(0, l)$ is said to be **totally symmetric** if $T_{a_1 \dots a_l} = T_{(a_1 \dots a_l)}$; T is said to be **totally antisymmetric** if $T_{a_1 \dots a_l} = T_{[a_1 \dots a_l]}$.

The concepts above (Definitions 1–3) can also be applied to tensors of type $(k, 0)$. For instance, T is said to be totally symmetric if $T^{a_1 \dots a_k} = T^{(a_1 \dots a_k)}$.

Remark 1 Any tensor of type $(0, 2)$ can be expressed as a summation of its symmetric part and antisymmetric part, i.e., $T_{ab} = T_{(ab)} + T_{[ab]}$. However, this is not true for tensors of type $(0, l)$ where $l > 2$. For example, $T_{abc} \neq T_{(abc)} + T_{[abc]}$, yet $T_{abc} = T_{(abc)} \Rightarrow T_{[abc]} = 0$ [see Theorem 2.6.2 (e)].

Theorem 2.6.1 (a) Suppose $T_{a_1 \dots a_l} = T_{(a_1 \dots a_l)}$, then

$$T_{a_1 \dots a_l} = T_{a_{\pi(1)} \dots a_{\pi(l)}} \quad (\text{where } \pi \text{ represents an arbitrary permutation}), \quad (2.6.15)$$

i.e., any term in the expansion of $T_{(a_1 \dots a_l)}$ ($l!$ terms in total) equals $T_{a_1 \dots a_l}$; for example,

$$T_{abc} = T_{(abc)} \Rightarrow T_{abc} = T_{acb} = T_{cab} = T_{cba} = T_{bca} = T_{bac}. \quad (2.6.16)$$

(b) Suppose $T_{a_1 \dots a_l} = T_{[a_1 \dots a_l]}$, then

$$T_{a_1 \dots a_l} = \delta_\pi T_{a_{\pi(1)} \dots a_{\pi(l)}}, \quad (2.6.17)$$

i.e., any even permutation term in the expansion of $T_{[a_1 \dots a_l]}$ equals $T_{a_1 \dots a_l}$, and any odd permutation term equals $-T_{a_1 \dots a_l}$; for example,

$$T_{abc} = T_{[abc]} \Rightarrow T_{abc} = -T_{acb} = T_{cab} = -T_{cba} = T_{bca} = -T_{bac}. \quad (2.6.18)$$

Similar conclusions also hold for the tensors of type $(k, 0)$ such that (the upper indices) are totally symmetric and totally antisymmetric.

Proof We only take the case of $l = 3$ as an example. For the other integer values of l , one can prove it in the same manner.

(a) From $T_{abc} = T_{(abc)}$ we have $T_{acb} = T_{(acb)}$ (the latter equation is nothing but a result of changing the abstract indices for both sides of the former one), and since $T_{(acb)} = T_{(abc)}$ (which is manifest from the definition of $T_{(abc)}$), we have $T_{acb} = T_{(acb)} = T_{abc}$. The other equalities of the right-hand side of (2.6.16) can be proved likewise.

(b) From $T_{abc} = T_{[abc]}$ we have $T_{acb} = T_{[acb]} = -T_{[abc]} = -T_{abc}$. The other equalities of the right-hand side of (2.6.18) can be proved likewise. \square

In the future we will often deal with the computations that involve parentheses and square brackets, and the theorem below will bring great convenience for many computations.

Theorem 2.6.2 (a) *The brackets are “contagious” in a contraction process, i.e.,*

$$T_{[a_1 \dots a_l]} S^{a_1 \dots a_l} = T_{[a_1 \dots a_l]} S^{[a_1 \dots a_l]} = T_{a_1 \dots a_l} S^{[a_1 \dots a_l]}, \quad (2.6.19)$$

and so are the parentheses.

(b) *One can arbitrarily add or delete one kind of bracket (parentheses or square brackets) inside a pair of the same kind of bracket; for example,*

$$T_{[[ab]c]} = T_{[abc]}, \quad \text{where } T_{[[ab]c]} \equiv \frac{1}{2}(T_{[abc]} - T_{[bac]}). \quad (2.6.20)$$

(c) *A pair of brackets inside a pair of the other kind of brackets yields zero; for example,*

$$T_{[(ab)c]} = 0, \quad T_{[a[bcd]]} = 0. \quad (2.6.21)$$

(d) *The contraction of different kinds of brackets yields zero; for example,*

$$T^{(abc)} S_{[abc]} = 0. \quad (2.6.22)$$

(e)

$$T_{a_1 \dots a_l} = T_{(a_1 \dots a_l)} \Rightarrow T_{[a_1 \dots a_l]} = 0, \quad (2.6.23)$$

$$T_{a_1 \dots a_l} = T_{[a_1 \dots a_l]} \Rightarrow T_{(a_1 \dots a_l)} = 0. \quad (2.6.24)$$

Similar conclusions also hold for the tensors of type $(k, 0)$ such that the upper indices are totally symmetric or totally antisymmetric.

Proof The proof of (a), (b), (c) are left as exercises. (d) is a corollary of (a) and (c), and (e) is a corollary of (c). \square

Exercises

- ~2.1. Show that the homeomorphism ψ_i^\pm defined in Example 2 of Sect. 2.1 satisfies the compatibility condition on all the overlap regions of O_i^\pm , which verifies that S^1 is indeed a 1-dimensional manifold.
- 2.2. Deduce that an n -dimensional vector space can be regarded as an n -dimensional trivial manifold.
- 2.3. Suppose X and Y are topological spaces, $f : X \rightarrow Y$ is a homeomorphism. If X is also a manifold, define a differential structure for Y such that $f : X \rightarrow Y$ is upgraded to a diffeomorphism.

- ~2.4. Suppose x, y are the natural coordinates of \mathbb{R}^2 , $C(t)$ is a curve whose parametric equations are $x = \cos t$ and $y = \sin t$, $t \in (0, \pi)$. If $p = C(\pi/3)$, write down the components of the tangent vector of the curve at p in the natural coordinate basis, and sketch this curve as well as this tangent vector.
- 2.5. Suppose the tangent vectors of two curves $C(t)$ and $C'(t) = C(2t_0 - t)$ at $C(t_0) = C'(t_0)$ are v and v' , respectively. Show that $v + v' = 0$.
- ~2.6. Suppose O is the coordinate patch of the coordinate system $\{x^\mu\}$, $p \in O$, $v \in V_p$, v^μ are the coordinate components of v . Regarding x^μ as a C^∞ function on O , show that $v^\mu = v(x^\mu)$. Hint: act both sides of $v = v^\nu X_\nu$ on a function x^μ .
- 2.7. Suppose M is a 2-dimensional manifold, (O, ψ) and (O', ψ') are two coordinate systems on M whose coordinates are x, y and x', y' , respectively, and the coordinate transformation on $O \cap O'$ is $x' = x$, $y' = y - \Omega x$ ($\Omega = \text{constant}$). Write down the expression for the expansion of $\partial/\partial x$ and $\partial/\partial y$ in terms of $\partial/\partial x'$ and $\partial/\partial y'$.
- ~2.8. (a) Show that $[u, v]$ in (2.2.9) pointwisely satisfies the two conditions in the definition of a vector (Definition 2 in Sect. 2.2), and thus is a vector field. (b) Suppose u, v, w are smooth vector fields on M . Show that

$$[[u, v], w] + [[w, u], v] + [[v, w], u] = 0 \quad (\text{this is called the } \mathbf{Jacobi identity}).$$

- ~2.9. Suppose r, φ are the polar coordinates on an open set (the coordinate patch) in \mathbb{R}^2 , x and y are natural coordinates.
- (a) Write down the expression for the expansion of the polar coordinate basis $\partial/\partial r$ and $\partial/\partial \varphi$ (as vector fields on the coordinate patch) in terms of $\partial/\partial x$ and $\partial/\partial y$.
- (b) Derive the expression for the expansion of a vector $[\partial/\partial r, \partial/\partial \varphi]$ in terms of $\partial/\partial x$ and $\partial/\partial y$.
- (c) Set $\hat{e}_r \equiv \partial/\partial r$, $\hat{e}_\varphi \equiv r^{-1} \partial/\partial \varphi$. Derive the expression for the expansion of $[\hat{e}_r, \hat{e}_\varphi]$ in terms of $\partial/\partial x$ and $\partial/\partial y$.
- ~2.10. Suppose u, v are vector fields on M . Show that the components of $[u, v]$ in any coordinate basis satisfy

$$[u, v]^\mu = u^\nu \frac{\partial v^\mu}{\partial x^\nu} - v^\nu \frac{\partial u^\mu}{\partial x^\nu}. \quad \text{Hint: use (2.2.3') and (2.2.3).}$$

- ~2.11. Suppose $\{e_\mu\}$ is a basis of V , and $\{e^{\mu*}\}$ is the dual basis, $v \in V$, $\omega \in V^*$. Show that

$$\omega = \omega(e_\mu)e^{\mu*}, \quad v = e^{\mu*}(v)e_\mu.$$

- ~2.12. Show that $\omega'_v = \frac{\partial x^\mu}{\partial x'^v} \omega_\mu$ (Theorem 2.3.4).
- ~2.13. Show that the map $v \rightarrow v^{**}$ defined by (2.3.5) is an isomorphism. Hint: one may use a conclusion of linear algebras, i.e., a one-to-one linear map between two vector spaces with the same dimension must be onto.

- ~2.14. Suppose $C_1^1 T$ and $(C_1^1 T)'$ are contractions of a tensor T of type (2, 1) defined in two different basis $\{e_\mu\}$ and $\{e'_\mu\}$. Show that $(C_1^1 T)' = C_1^1 T$.
- *2.15. Suppose g is a metric of V . Show that $g : V \rightarrow V^*$ is an isomorphism (see the hint for Exercise 2.13).
- ~2.16. Show that the arc length of a curve does not depend on the parametrization.
- 2.17. Suppose $\{x, y\}$ is a Cartesian coordinate system of 2-dimensional Euclidean space. Show that $\{x', y'\}$ defined by (2.5.14) is also a Cartesian system.
- 2.18. Suppose $\{t, x\}$ is a Lorentzian coordinate system of 2-dimensional Minkowski space. Show that $\{t', x'\}$ defined by (2.5.20) is also a Lorentzian system.
- ~2.19. (a) Using the tensor transformation law, derive all the components $g'_{\mu\nu}$ of the 3-dimensional Euclidean metric in a spherical coordinate system. (b) Given the expression for the line element of the 4-dimensional Minkowski metric in a Lorentzian system $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$, derive all the components of g and its inverse g^{-1} in a new coordinate system $\{t', x', y', z'\}$, denoted by $g'_{\mu\nu}$ and $g'^{\mu\nu}$. This new coordinate system is defined as follows:

$$\begin{aligned} t' &= t, \quad z' = z, \quad x' = (x^2 + y^2)^{1/2} \cos(\varphi - \omega t), \\ y' &= (x^2 + y^2)^{1/2} \sin(\varphi - \omega t), \quad \omega = \text{constant}, \end{aligned}$$

where φ satisfies $\cos \varphi = y(x^2 + y^2)^{-1/2}$, $\sin \varphi = x(x^2 + y^2)^{-1/2}$. Hint: first derive $g'^{\mu\nu}$ and then derive $g'_{\mu\nu}$.

- ~2.20. Show that the lengths of the spherical coordinate basis vectors $\partial/\partial r$, $\partial/\partial\theta$, $\partial/\partial\varphi$ in 3-dimensional Euclidean space are 1, r and $r \sin \theta$.
- ~2.21. Using the abstract index notation, show that $T'^\mu{}_\nu = \frac{\partial x'^\mu}{\partial x^\rho} \frac{\partial x^\sigma}{\partial x'^\nu} T^\rho{}_\sigma$.
- 2.22. Using g and g' to represent the two $n \times n$ matrices constituted by the components $g_{\mu\nu}$ and $g'_{\mu\nu}$ of g_{ab} in coordinate systems $\{x^\mu\}$ and $\{x'^\mu\}$, respectively, show that $g' = |\partial x^\rho / \partial x'^\sigma|^2 g$, where $|\partial x^\rho / \partial x'^\sigma|$ is the Jacobian determinant of the coordinate transformation $\{x^\mu\} \mapsto \{x'^\mu\}$, i.e., the $n \times n$ determinant constituted by $\partial x^\rho / \partial x'^\sigma$. NB: This exercise indicates that the determinant of a metric is not an invariant under a coordinate transformation. Hint: take the determinant of the equality $g'_{\rho\sigma} = (\partial x^\mu / \partial x'^\rho)(\partial x^\nu / \partial x'^\sigma)g_{\mu\nu}$.
- ~2.23. Suppose $\{x^\mu\}$ is an arbitrary local coordinate system on a manifold. Determine whether each of the following equation is true or false:
- (1) $(\partial/\partial x^\mu)^a (\partial/\partial x^\nu)_a = g_{\mu\nu}$, where $(\partial/\partial x^\nu)_a \equiv g_{ab}(\partial/\partial x^\nu)^b$;
 - (2) $(dx^\mu)^a (dx^\nu)_a = g^{\mu\nu}$, where $(dx^\mu)^a \equiv g^{ab}(dx^\mu)_b$;
 - (3) $(\partial/\partial x^\mu)_a = (dx^\mu)_a$;
 - (4) $(dx^\mu)^a = (\partial/\partial x^\mu)^a$;
 - (5) $v^\mu \omega_\mu = v_\mu \omega^\mu$;
 - (6) $g_{\mu\nu} T^{\nu\rho} S_\rho{}^\sigma = T_{\mu\rho} S^{\rho\sigma}$;
 - (7) $v^a u^b = v^b u^a$;
 - (8) $v^a u^b = u^b v^a$.

- 2.24. Suppose T_{ab} is a tensor of type $(0, 2)$ on a vector space V . Show that $T_{ab}v^a v^b = 0, \forall v^a \in V \Rightarrow T_{ab} = T_{[ab]}$. Hint: express v^a as the sum of two arbitrary vectors u^a and w^a .
- 2.25. Show that $T_{abcd} = T_{a[bc]d} = T_{ab[cd]} \Rightarrow T_{abcd} = T_{a[bcd]}$.
- Remark** (1) The above claim has the following generalization:

$$T_{\dots a \dots b \dots c \dots} = T_{\dots [a \dots b] \dots c \dots} = T_{\dots a \dots [b \dots c] \dots} \Rightarrow T_{\dots a \dots b \dots c \dots} = T_{\dots [a \dots b \dots c] \dots}.$$

The premise above only contains two equal signs, the key point is that the index b from both $T_{\dots [a \dots b] \dots c \dots}$ and $T_{\dots a \dots [b \dots c] \dots}$ are inside the square brackets.

(2) Both the original and generalized claims will still hold when changing the square brackets in the premise and conclusion to parentheses.

References

- Hawking, S. W. and Ellis, G. F. R. (1973), *The Large Scale Structure of Space-Time*, Cambridge University Press, Cambridge.
- Kline, M. (1980), *Mathematics: The Loss of Certainty*, Oxford University Press, New York.
- Sachs, R. K. and Wu, H. (1977), *General Relativity for Mathematicians*, Springer-Verlag, New York.
- Schutz, B. F. (1980), *Geometrical Methods of Mathematical Physics*, Cambridge University Press, Cambridge.
- Spivak, M. (1970), *A Comprehensive Introduction to Differential Geometry*, Vol. I, II, Publish or Perish INC, Berkeley.
- Straumann, N. (1984), *General Relativity and Relativistic Astrophysics*, Springer-Verlag, Berlin.
- Wald, R. M. (1984), *General Relativity*, The University of Chicago Press, Chicago.

Chapter 3

The Riemann (Intrinsic) Curvature Tensor



3.1 Derivative Operators

In Euclidean space there is a familiar derivative operator $\vec{\nabla}$, the action of which on, for example, a function (scalar field) f yields a vector field $\vec{\nabla}f$ (gradient) and on a vector field \vec{v} (with contraction) it yields a scalar field $\vec{\nabla} \cdot \vec{v}$ (divergence). Since there exists a Euclidean metric δ_{ab} , a vector v^a can be naturally identified with a dual vector $v_a = \delta_{ab}v^b$. Now we want to generalize $\vec{\nabla}$ to an arbitrary manifold that may not have a metric, so we need to distinguish vectors and dual vectors. It has been shown that $\vec{\nabla}$ behaves more like a dual vector after being generalized, and hence should be denoted by ∇_a . Actually, ∇ itself is an operator, which is neither a vector nor a dual vector; by regarding ∇ as a dual vector, we mean that the result of it acting on a function f is a dual vector $\nabla_a f$. More generally, the result of ∇ acting on a tensor field of type (k, l) is a tensor field of type $(k, l + 1)$. Therefore, we have the following definition:

Definition 1 Use $\mathcal{F}_M(k, l)$ to represent the collection of all C^∞ tensor fields of type (k, l) on a manifold M . [A function f can be viewed as a tensor field of type $(0, 0)$ (scalar field), and hence $\mathcal{F}_M(0, 0) \equiv \mathcal{F}_M$.] A map $\nabla : \mathcal{F}_M(k, l) \rightarrow \mathcal{F}_M(k, l + 1)$ is called a **derivative operator**¹ on M if it satisfies the following conditions:

(a) Linearity:

$$\nabla_a(\alpha T^{b_1 \dots b_k}_{c_1 \dots c_l} + \beta S^{b_1 \dots b_k}_{c_1 \dots c_l}) = \alpha \nabla_a T^{b_1 \dots b_k}_{c_1 \dots c_l} + \beta \nabla_a S^{b_1 \dots b_k}_{c_1 \dots c_l}$$

$$\forall T^{b_1 \dots b_k}_{c_1 \dots c_l}, S^{b_1 \dots b_k}_{c_1 \dots c_l} \in \mathcal{F}_M(k, l), \quad \alpha, \beta \in \mathbb{R};$$

(b) Leibniz rule:

$$\nabla_a(T^{b_1 \dots b_k}_{c_1 \dots c_l} S^{d_1 \dots d_{k'}}_{e_1 \dots e_{l'}}) = T^{b_1 \dots b_k}_{c_1 \dots c_l} \nabla_a S^{d_1 \dots d_{k'}}_{e_1 \dots e_{l'}} + S^{d_1 \dots d_{k'}}_{e_1 \dots e_{l'}} \nabla_a T^{b_1 \dots b_k}_{c_1 \dots c_l}$$

¹ $\mathcal{F}(k, l)$ can be relaxed to the collection of all C^1 tensor fields of type (k, l) ; that is, ∇_a can act on an arbitrary tensor field of class C^1 .

$$\forall T^{b_1 \dots b_k}_{c_1 \dots c_l} \in \mathcal{F}_M(k, l), S^{d_1 \dots d_{k'}}_{e_1 \dots e_{l'}} \in \mathcal{F}_M(k', l');$$

- (c) Commutativity with contraction;
(d) $v(f) = v^a \nabla_a f$, $\forall f \in \mathcal{F}_M, v \in \mathcal{F}_M(1, 0)$.

Remark 1 (1) Condition (c) can also be expressed as $\nabla \circ C = C \circ \nabla$, where C stands for contraction. In the future, we will often write equations like

$$\nabla_a(v^b \omega_b) = v^b \nabla_a \omega_b + \omega_b \nabla_a v^b,$$

which requires condition (c) since the derivation of this equation reads

$$\begin{aligned} \nabla_a(v^b \omega_b) &= \nabla_a[C(v^b \omega_c)] = C_2^1[\nabla_a(v^b \omega_c)] \\ &= C_2^1(v^b \nabla_a \omega_c) + C_2^1[(\nabla_a v^b) \omega_c] = v^b \nabla_a \omega_b + \omega_b \nabla_a v^b, \end{aligned}$$

where we used condition (c) in the second step (see Sect. 2.4 for a refresher on the operation of C).

(2) The function $v(f)$ on the left-hand side of condition (d) should not be denoted by $v^a(f)$ since it may be easily mistaken for a vector field. This is one of the few cases where we should but we do not put on an abstract index. To understand condition (d), one can use $\vec{\nabla}$ in Euclidean space as an example. Suppose v^a is a vector field in Euclidean space whose expansion in the Cartesian coordinates is

$$v^a = v^1(\partial/\partial x)^a + v^2(\partial/\partial y)^a + v^3(\partial/\partial z)^a,$$

then the action of it on a function f can be expressed as

$$v(f) = v^1(\partial f/\partial x) + v^2(\partial f/\partial y) + v^3(\partial f/\partial z) = \vec{v} \cdot \vec{\nabla} f = v^a \nabla_a f.$$

Thus, condition (d) is a generalization of this property to a general manifold.

(3) Suppose ∇_a is an arbitrary derivative operator, then from condition (d) it is easy to show that (exercise)

$$\nabla_a f = (df)_a, \quad \forall f \in \mathcal{F}_M, \tag{3.1.1}$$

where $(df)_a$ is the abstract index expression for a dual vector field df generated by a function f [see (2.3.7)].

(4) In general relativity, a derivative operator also satisfies $\nabla_a \nabla_b f = \nabla_b \nabla_a f$, $\forall f \in \mathcal{F}_M$. From Definition 1 in Sect. 2.6 we can see that this is essentially the abstract index expression for

$$(\nabla \nabla f)(u, v) = (\nabla \nabla f)(v, u), \quad \forall u, v \in \mathcal{F}_M(1, 0),$$

which means $\nabla\nabla f$ is a symmetric tensor of type $(0, 2)$. A derivative operator that satisfies this additional condition is called a **torsion-free derivative operator**. Unless stated otherwise, all ∇_a in this text will stand for torsion-free derivative operators.

[Optional Reading 3.1.1]

This optional reading has the same spirit as Optional Reading 2.2.1. For the sake of conciseness, here we abbreviate a tensor field $T^{b_1 \dots b_k}_{c_1 \dots c_l}$ as T .

Theorem 3.1.1 Suppose $T_1, T_2 \in \mathcal{F}_M(k, l)$ are equal in a neighborhood N of $p \in M$, i.e., $T_1|_N = T_2|_N$, then $\nabla_a T_1|_p = \nabla_a T_2|_p$.

Proof The proof is similar to that of Theorem 2.2.1, and should be carried out by the reader. \square

Remark 2 Suppose a tensor field T is only defined on a neighborhood U of $p \in M$ ($\neq U$), i.e., $T \in \mathcal{F}_U(k, l)$, $T \notin \mathcal{F}_M(k, l)$. According to Definition 1, ∇_a can only act on a tensor field on M , and so $\nabla_a T$ is meaningless. However, one can always find a $\tilde{T} \in \mathcal{F}_M(k, l)$ and a neighborhood $N \subset U$ of p such that $\tilde{T}|_N = T|_N$, and thus one can define $\nabla_a T$ as $\nabla_a \tilde{T}$. Although for the same T there are infinitely many \tilde{T} that satisfies the requirement above, Theorem 3.1.1 guarantees that $\nabla_a \tilde{T}$ are the same for all \tilde{T} . Thus, it is legal to define $\nabla_a T$ as $\nabla_a \tilde{T}$. Therefore, we say that ∇_a is a **local operator**, the action of which on T has a value at p that only depends on the behavior of T on a neighborhood of p (no matter how “small” it is). The reader may already be familiar with the similar property of the derivative of a function in calculus.

[The End of Optional Reading 3.1.1]

For any manifold, there always exists a derivative operator that satisfies Definition 1 [see Theorem 1.1 in Chap. 4 of Chern et al. (1999)]. In fact, derivative operators on a manifold not only exist, but also they are numerous. Now we will discuss how many there can be. From (3.1.1) we know that two different derivative operators ∇_a and $\tilde{\nabla}_a$ acting on the same function give the same result, i.e.,

$$\nabla_a f = \tilde{\nabla}_a f = (df)_a, \quad \forall f \in \mathcal{F}_M. \quad (3.1.2)$$

Thus, the difference between ∇_a and $\tilde{\nabla}_a$ can only be manifested by the action on a tensor field not of type $(0, 0)$. First we discuss the action on a tensor field of type $(0, 1)$ (a dual vector field). Suppose a dual vector $\mu_b \in V_p^*$ is given at a point $p \in M$, and consider two arbitrary dual vector fields $\omega_b, \omega'_b \in \mathcal{F}_M(0, 1)$ on M that satisfy $\omega'_b|_p = \omega_b|_p = \mu_b$ (ω_b and ω'_b are called two extensions of μ_b on M). Suppose ∇_a is a derivative operator on M , then $\nabla_a \omega'_b|_p$ and $\nabla_a \omega_b|_p$ are not the same in general. This is similar to the fact that two functions $f(x)$ and $f'(x)$ that have the same value at x_0 [i.e., $f'(x_0) = f(x_0)$] are not assured to have $(df'/dx)|_{x_0} = (df/dx)|_{x_0}$. However, we are about to show that for any two derivative operators ∇_a and $\tilde{\nabla}_a$ on M , as long as $\omega'_b|_p = \omega_b|_p$, we have

$$[(\tilde{\nabla}_a - \nabla_a)\omega'_b]_p = [(\tilde{\nabla}_a - \nabla_a)\omega_b]_p,$$

where $(\tilde{\nabla}_a - \nabla_a)\omega_b$ is short for $\tilde{\nabla}_a \omega_b - \nabla_a \omega_b$.

Theorem 3.1.2 Suppose $p \in M$ and $\omega_b, \omega'_b \in \mathcal{F}_M(0, 1)$ satisfy $\omega'_b|_p = \omega_b|_p$, then

$$[(\tilde{\nabla}_a - \nabla_a)\omega'_b]_p = [(\tilde{\nabla}_a - \nabla_a)\omega_b]_p. \quad (3.1.3)$$

Proof Alternately, this equation can be rearranged as

$$[\nabla_a(\omega'_b - \omega_b)]_p = [\tilde{\nabla}_a(\omega'_b - \omega_b)]_p. \quad (3.1.4)$$

Suppose $\Omega_b \equiv \omega'_b - \omega_b$. Choose a coordinate system $\{x^\mu\}$ such that its coordinate patch includes p , then $\omega'_b|_p = \omega_b|_p$ leads to $\Omega_\mu(p) = 0$, where Ω_μ are the coordinate components of Ω_b . Hence, at p we have

$$\begin{aligned} [\nabla_a(\omega'_b - \omega_b)]_p &= [\nabla_a\Omega_b]_p = \{\nabla_a[\Omega_\mu(dx^\mu)_b]\}|_p \\ &= \Omega_\mu(p)[\nabla_a(dx^\mu)_b]_p + [(dx^\mu)_b\nabla_a\Omega_\mu]_p = [(dx^\mu)_b\nabla_a\Omega_\mu]_p, \end{aligned}$$

Similarly we have $[\tilde{\nabla}_a(\omega'_b - \omega_b)]_p = [(dx^\mu)_b\tilde{\nabla}_a\Omega_\mu]_p$. From (3.1.2) we know that $[\nabla_a\Omega_\mu]_p = [\tilde{\nabla}_a\Omega_\mu]_p$, which completes the proof. \square

Although $[\nabla_a\omega_b]_p$ and $[\tilde{\nabla}_a\omega_b]_p$ depend on the value of ω_b in a neighborhood of p , Theorem 3.1.2 indicates that $[(\tilde{\nabla}_a - \nabla_a)\omega_b]_p$ only depends on the value of ω_b at p . This means that $(\tilde{\nabla}_a - \nabla_a)$ is a linear map that turns $\omega_b|_p$, a dual vector at p , into $[(\tilde{\nabla}_a - \nabla_a)\omega_b]_p$, which implies that $[(\tilde{\nabla}_a - \nabla_a)\omega_b]_p$ is a tensor of type $(0, 2)$ at p . (For a given dual vector μ_b at p , we choose any dual vector field ω_b such that $\omega_b|_p = \mu_b$, then $[(\tilde{\nabla}_a - \nabla_a)\omega_b]_p$ is the image of μ_b under this linear map.) Therefore, $(\tilde{\nabla}_a - \nabla_a)$ at p corresponds to a tensor C^c_{ab} of type $(1, 2)$, which satisfies

$$[(\tilde{\nabla}_a - \nabla_a)\omega_b]_p = C^c_{ab}\omega_c|_p. \quad (3.1.5)$$

Since p is chosen arbitrarily, the difference between two derivative operators ∇_a and $\tilde{\nabla}_a$ on M is manifested by a tensor C^c_{ab} of type $(1, 2)$; that is:

Theorem 3.1.3

$$\nabla_a\omega_b = \tilde{\nabla}_a\omega_b - C^c_{ab}\omega_c, \quad \forall \omega_b \in \mathcal{F}(0, 1). \quad (3.1.6)$$

∇_a being torsion free will give rise to the following symmetry of the tensor field C^c_{ab} :

Theorem 3.1.4 $C^c_{ab} = C^c_{ba}$.

Proof Let $\omega_b = \nabla_b f = \tilde{\nabla}_b f$ [(3.1.2)] where $f \in \mathcal{F}_M$, then (3.1.6) will yield $\nabla_a\nabla_b f = \tilde{\nabla}_a\tilde{\nabla}_b f - C^c_{ab}\nabla_c f$. Switching the indices a and b , we get $\nabla_b\nabla_a f = \tilde{\nabla}_b\tilde{\nabla}_a f - C^c_{ba}\nabla_c f$. Subtracting these two equations and noticing the torsion-free condition, we have $C^c_{ab}\nabla_c f = C^c_{ba}\nabla_c f$. Let $T^c_{ab} \equiv C^c_{ab} - C^c_{ba}$, then $\forall f \in \mathcal{F}_M$ we have $T^c_{ab}\nabla_c f = 0$, and hence the components of T^c_{ab} in an arbitrary coordinate basis

are $T^\sigma_{\mu\nu} = T^c_{ab}(\mathrm{d}x^\sigma)_c(\partial/\partial x^\mu)^a(\partial/\partial x^\nu)^b = 0$ [where the second step is because $T^c_{ab}(\mathrm{d}x^\sigma)_c = T^c_{ab}\nabla_c x^\sigma = 0$ (regard x^σ as f)]. Therefore, $T^c_{ab} = 0$. \square

Theorem 3.1.5

$$\nabla_a v^b = \tilde{\nabla}_a v^b + C^b_{ac} v^c \quad \forall v^b \in \mathcal{F}_M(1, 0). \quad (3.1.7)$$

Proof Suppose ω_b is an arbitrary dual vector field on M , then

$$\nabla_a(\omega_b v^b) = \omega_b \nabla_a v^b + v^b \nabla_a \omega_b = \omega_b \nabla_a v^b + v^b (\tilde{\nabla}_a \omega_b - C^c_{ab} \omega_c),$$

where we used (3.1.6) in the last step. On the other hand, $\tilde{\nabla}_a(\omega_b v^b) = \omega_b \tilde{\nabla}_a v^b + v^b \tilde{\nabla}_a \omega_b$. Since $\omega_b v^b$ is a scalar field, it follows from (3.1.2) that $\nabla_a(\omega_b v^b) = \tilde{\nabla}_a(\omega_b v^b)$, and hence the right-hand sides of the two equations above are equal. Therefore, we obtain

$$\omega_b \nabla_a v^b = \omega_b \tilde{\nabla}_a v^b + C^c_{ab} v^b \omega_c = \omega_b \tilde{\nabla}_a v^b + C^b_{ac} v^c \omega_b, \quad \forall \omega_b \in \mathcal{F}_M(0, 1),$$

and thus we have (3.1.7). \square

By a similar analysis, one can also show that the difference between the result of ∇_a and $\tilde{\nabla}_a$ acting on a tensor field $T^{b_1 \dots b_k}_{c_1 \dots c_l}$ of type (k, l) , i.e., $\nabla_a T^{b_1 \dots b_k}_{c_1 \dots c_l} - \tilde{\nabla}_a T^{b_1 \dots b_k}_{c_1 \dots c_l}$, can be expressed in $k + l$ terms, each of which has a C^c_{ab} . In front of each term there is a + sign if it contracts with an upper index of T , and a - sign if it contracts with a lower index of T ; for example,

$$\nabla_a T^b{}_c = \tilde{\nabla}_a T^b{}_c + C^b{}_{ad} T^d{}_c - C^d{}_{ac} T^b{}_d,$$

and the general form is given in the following theorem:

Theorem 3.1.6

$$\begin{aligned} \nabla_a T^{b_1 \dots b_k}_{c_1 \dots c_l} &= \tilde{\nabla}_a T^{b_1 \dots b_k}_{c_1 \dots c_l} + \sum_i C^{b_i}{}_{ad} T^{b_1 \dots d \dots b_k}_{c_1 \dots c_l} - \sum_j C^d{}_{ac_j} T^{b_1 \dots b_k}_{c_1 \dots d \dots c_l} \\ &\quad \forall T \in \mathcal{F}_M(k, l). \end{aligned} \quad (3.1.8)$$

Proof Exercise. \square

Theorem 3.1.6 indicates that the difference between two arbitrary derivative operators is only manifested by a tensor field C^c_{ab} . Conversely, it is not difficult to verify that given an arbitrary derivative operator $\tilde{\nabla}_a$ and a smooth tensor field C^c_{ab} with symmetric lower indices, ∇_a defined by (3.1.8) satisfies all of the conditions in Definition 1, and thus this ∇_a is also a derivative operator. Therefore, there exists numerous derivative operators on a manifold as long as there is one. A manifold with a chosen derivative operator can be denoted by (M, ∇_a) , and this combination has

more structure than M itself (∇_a provides additional structure); for instance, we can now talk about the parallel transport of a vector along a curve (see Sect. 3.2) and the curvature of (M, ∇_a) (see Sect. 3.4).

Suppose $\{x^\mu\}$ is a coordinate system of M , the coordinate basis and dual basis of which are $\{(\partial/\partial x^\mu)^a\}$ and $\{(dx^\mu)_a\}$. Define a map $\partial_a : \mathcal{F}_O(k, l) \rightarrow \mathcal{F}_O(k, l+1)$ on the coordinate patch of O as follows [we only write down the case $T^b_c \in \mathcal{F}_O(1, 1)$ as an example]:

$$\partial_a T^b_c := (dx^\mu)_a (\partial/\partial x^\nu)^b (dx^\sigma)_c \partial_\mu T^\nu_\sigma , \quad (3.1.9)$$

where T^ν_σ are the components of T^b_c in this coordinate system, and ∂_μ is short for $\partial/\partial x^\mu$, the partial derivative with respect to a coordinate x^μ . It is not difficult to verify that ∂_a satisfies all of the conditions in Definition 1 plus the torsion-free condition, and thus ∂_a is a torsion-free derivative operator on O . This is a derivative operator that by definition depends on the coordinate system, and it is only defined in the coordinate patch of this coordinate system, called the **ordinary derivative operator** of this coordinate system. Equation (3.1.9) indicates that $\partial_\mu T^\nu_\sigma$ are the components of $\partial_a T^b_c$ in this coordinate system, and therefore the definition of ∂_a can also be formulated as: the coordinate components of the ordinary derivative $\partial_a T^{b_1 \dots b_k}_{c_1 \dots c_l}$ of a tensor field $T^{b_1 \dots b_k}_{c_1 \dots c_l}$ are equal to $\partial(T^{v_1 \dots v_k}_{\sigma_1 \dots \sigma_l})/\partial x^\mu$, the derivatives of the coordinate components of this tensor field with respect to the coordinates. Thus, we can easily see that:

(1) ∂_a of any coordinate system acting on a coordinate basis vector and a dual coordinate basis vector of this system yields zero, i.e.,

$$\partial_a (\partial/\partial x^\nu)^b = 0 , \quad \partial_a (dx^\mu)_b = 0 . \quad (3.1.10)$$

(2) ∂_a satisfies a much stronger condition than the torsion-free condition, i.e.,

$$\partial_a \partial_b T^{\dots} \dots = \partial_b \partial_a T^{\dots} \dots , \quad \text{or } \partial_{[a} \partial_{b]} T^{\dots} \dots = 0 ,$$

where $T^{\dots} \dots$ is a tensor field of any type.

Although ∂_a can be viewed as a special case of ∇_a , the definition of it depends on the coordinate system. We call those ∇_a that are independent of a coordinate system (and any other externally imposed factor) **covariant derivative operators**, in which ∂_a is not included.

Definition 2 Suppose ∂_a is an ordinary derivative operator of a given coordinate system on (M, ∇_a) , then the tensor field C^c_{ab} that manifests the difference between ∇_a and ∂_a [regard ∂_a as $\tilde{\nabla}_a$ in (3.1.6)] is called the **Christoffel symbol** of ∇_a in this coordinate system, denoted by Γ^c_{ab} .

Remark 3 Normally, textbooks may emphasize that Christoffel symbols are not tensors, while this text and some other books [e.g., Wald (1984)] call it a tensor instead. There is no substantial conflict between them, it is just the subtle difference in the definition of a Christoffel symbol. In the books that use the component index notation, a Christoffel symbol is defined as an array of numbers which do not obey the

tensor transformation law under a coordinate transformation, and hence do not constitute a tensor. From the very beginning, however, we define a Christoffel symbol as a tensor, which is a multilinear map, but since it corresponds to ∂_a which depends on the coordinate system, a Christoffel symbol is a tensor associated with the coordinate system (the tensor *itself* will change under a coordinate transformation). Suppose ∇_a is a derivative operator assigned on M , $\{x^\mu\}$ and $\{x'^\mu\}$ are two coordinate systems on M , the intersection of their coordinate patches is U , and the Christoffel symbols of ∇_a in these two systems are Γ^c_{ab} and $\tilde{\Gamma}^c_{ab}$, respectively. As tensors, they can be expressed as components (in U) using the $\{x^\mu\}$ system or the $\{x'^\mu\}$ system. Suppose the components of Γ^c_{ab} in the $\{x^\mu\}$ and $\{x'^\mu\}$ systems are $\{\Gamma^\sigma_{\mu\nu}\}$ and $\{\tilde{\Gamma}^\sigma_{\mu\nu}\}$ (these two arrays of numbers certainly satisfy the tensor transformation law), and the components of $\tilde{\Gamma}^c_{ab}$ in the $\{x^\mu\}$ and $\{x'^\mu\}$ systems are $\{\tilde{\Gamma}^\sigma_{\mu\nu}\}$ and $\{\tilde{\Gamma}'^\sigma_{\mu\nu}\}$ (which also satisfy the tensor transformation law); however, $\{\Gamma^\sigma_{\mu\nu}\}$ and $\{\tilde{\Gamma}'^\sigma_{\mu\nu}\}$ do not satisfy the tensor transformation law. Nevertheless, textbooks normally just define $\{\Gamma^\sigma_{\mu\nu}\}$ and $\{\tilde{\Gamma}'^\sigma_{\mu\nu}\}$ to be the Christoffel symbols in the coordinate systems $\{x^\mu\}$ and $\{x'^\mu\}$, respectively, and therefore there is no doubt that they do not constitute a tensor. It is right for those books to emphasize “a Christoffel symbol is not a tensor”, but we instead emphasize that “a Christoffel symbol is a tensor associated with a coordinate system”. The reader may ask: why do you have to describe a Christoffel symbol as a tensor? The answer is: as long as we use the abstract index notation and follow the above reasoning (including the elegant argument from the “multifaceted view of tensor”), surely we need to admit that C^c_{ab} is a tensor that reflects the difference between ∇_a and $\tilde{\nabla}_a$. Under the premise that a derivative operator has been assigned to M , for a given coordinate system there is a derivative ∂_a , and if we regard ∂_a as $\tilde{\nabla}_a$, then C^c_{ab} (which is now denoted by Γ^c_{ab}) is, of course, a tensor. It would be a slap in the face if we do not admit that Γ^c_{ab} is a tensor. However, at the same time, we should emphasize that Γ^c_{ab} is a tensor associated with a coordinate system. (There are as many ∂_a , and thus as many Γ^c_{ab} , as there are coordinate systems). This emphasis is essentially the same as the emphasis in many books that say “a Christoffel symbol is not a tensor”. They are just two ways of wording the same issue. What is important is not how it is worded but the substance of it, i.e., we should keep in mind that it does not satisfy the tensor transformation law between $\{\Gamma^\sigma_{\mu\nu}\}$ and $\{\tilde{\Gamma}'^\sigma_{\mu\nu}\}$.

Similarly, suppose v^b is a vector field, then $\partial_a v^b$ is also a tensor field associated with the coordinate system. Expand $\partial_a v^b$ in the coordinate system associated with ∂_a :

$$\partial_a v^b = (dx^\mu)_a (\partial/\partial x^\nu)^b v^\nu,_\mu ,$$

where $v^\nu,_\mu \equiv \partial_\mu v^\nu \equiv \partial v^\nu / \partial x^\mu$ (the comma stands for the partial derivative). Again, textbooks often emphasize that $v^\nu,_\mu$ does not constitute a tensor, while we say that $\partial_a v^b$ is a tensor field associated with the coordinate system; they are also just two ways of wording the same issue. More specifically, suppose ∂_a and ∂'_a are the ordinary derivative operators of two coordinate systems $\{x^\mu\}$ and $\{x'^\mu\}$, respectively, then usually $\partial_a v^b \neq \partial'_a v^b$ (that is why $\partial_a v^b$ is a tensor field associated with the coordinate system). If we expand $\partial_a v^b$ and $\partial'_a v^b$ in terms of their own coordinate basis:

$$\partial_a v^b = (\mathrm{d}x^\mu)_a (\partial/\partial x^\nu)^b v^\nu{}_{,\mu}, \quad \partial'_a v^b = (\mathrm{d}x'^\mu)_a (\partial/\partial x'^\nu)^b v'^\nu{}_{,\mu},$$

where $v^\nu{}_{,\mu} \equiv \partial v^\nu / \partial x^\mu$, then $\partial_a v^b \neq \partial'_a v^b$ makes it so the tensor transformation law is not generally satisfied between $v^\nu{}_{,\mu}$ and $v'^\nu{}_{,\mu}$ (this can also be verified directly, see Exercise 3.2). That is why textbooks usually say that $v^\nu{}_{,\mu}$ is not a tensor. As for $\nabla_a v^b$, it is a tensor independent of a coordinate system whose components in a coordinate system are usually denoted by $v^\nu{}_{;\mu}$, i.e., $\nabla_a v^b = v^\nu{}_{;\mu} (\mathrm{d}x^\mu)_a (\partial/\partial x^\nu)^b$. Since $\nabla_a v^b$ is independent of a coordinate system, and $v^\nu{}_{;\mu}$ satisfies the tensor transformation rule; thus, textbooks usually say that it is a tensor (actually, the components of a tensor), and call it the **covariant derivative** of v^ν (actually, the coordinate components of the covariant derivative $\nabla_a v^b$). Similarly, $\omega_{v;\mu}$, the coordinate components of $\nabla_a \omega_b$, are also called the covariant derivative of ω_v .

Theorem 3.1.7

$$v^\nu{}_{;\mu} = v^\nu{}_{,\mu} + \Gamma^\nu{}_{\mu\sigma} v^\sigma, \quad \omega_{v;\mu} = \omega_{v,\mu} - \Gamma^\sigma{}_{\mu\nu} \omega_\sigma, \quad (3.1.11)$$

where v^ν and ω_v are the components of any vector field and dual vector field in an arbitrary coordinate basis, $\Gamma^\nu{}_{\mu\sigma}$ are the components of the Christoffel symbol of this system in this basis. (Many books may say “ $\Gamma^\nu{}_{\mu\sigma}$ is the Christoffel symbol of this system”; later on, we will also say this for simplicity.)

Proof Exercise 3.3. □

Theorem 3.1.8 Condition (c) in Definition 1 is equivalent to

$$\nabla_a \delta^b{}_c = 0, \quad (3.1.12)$$

where $\delta^b{}_c$ is a tensor field of type $(1, 1)$, whose definition at each point $p \in M$ is $\delta^b{}_c v^c = v^b, \forall v^c \in V_p$.

Proof [Optional Reading]

(A) Suppose ∇_a satisfies all of the conditions in Definition 1, we would like to show that it satisfies (3.1.12). $\forall v^b \in \mathcal{F}_M(1, 0)$ we have

$$\begin{aligned} \nabla_a v^b &= \nabla_a (\delta^b{}_c v^c) = \nabla_a [C(\delta^b{}_c v^d)] = C[\nabla_a (\delta^b{}_c v^d)] \\ &= C(v^d \nabla_a \delta^b{}_c + \delta^b{}_c \nabla_a v^d) = v^c \nabla_a \delta^b{}_c + \delta^b{}_c \nabla_a v^c = v^c \nabla_a \delta^b{}_c + \nabla_a v^b, \end{aligned}$$

where C stands for the contraction of indices c and d ; in the third equality we used condition (c) and in the last step we used $\delta^b{}_c T_a{}^c = T_a{}^b \forall T_a{}^c$. The above equation indicates that $v^c \nabla_a \delta^b{}_c = 0 \forall v^c \in \mathcal{F}_M(1, 0)$, and therefore $\nabla_a \delta^b{}_c = 0$.

(B) Suppose $\tilde{\nabla}_a$ satisfies conditions (a), (b), (d) in Definition 1 and (3.1.12). We would like to show that it also satisfies condition (c). For this propose, suppose ∇_a satisfies all of the conditions in Definition 1. Since the proof of Theorem 3.1.2 does not need condition (c), (3.1.6) is satisfied. We cannot use Theorem 3.1.4 directly since the proof of it needs condition (c); however, one can still prove it from the properties of ∇_a and $\tilde{\nabla}_a$ (motivated readers may try it as a challenging exercise), and therefore we have (3.1.8). From this, using the fact that ∇_a satisfies (c) we can show that $\tilde{\nabla}_a$ satisfies condition (c). □

The commutator $[u, v]^a$ of two vector fields on M does not require M to have any additional structure [see (2.2.9)]; however, the inconvenience of this equation is that it cannot be isolated by the object it acts on (a scalar field f). Now, after we have the concept of a derivative operator, we can write the explicit expression of a commutator of vector fields $[u, v]^a$ by means of an arbitrary torsion-free derivative operator, as shown in the following theorem:

Theorem 3.1.9

$$[u, v]^a = u^b \nabla_b v^a - v^b \nabla_b u^a, \quad (3.1.13)$$

where ∇_b is an arbitrary torsion-free derivative operator.

Proof $\forall f \in \mathcal{F}_M$ we have

$$\begin{aligned} [u, v](f) &= u(v(f)) - v(u(f)) = u^b \nabla_b (v^a \nabla_a f) - v^b \nabla_b (u^a \nabla_a f) \\ &= u^b (\nabla_b v^a) \nabla_a f + v^a u^b \nabla_b \nabla_a f - v^b (\nabla_b u^a) \nabla_a f - u^a v^b \nabla_b \nabla_a f \\ &= (u^b \nabla_b v^a - v^b \nabla_b u^a) \nabla_a f, \end{aligned}$$

where in the second step we used condition (d) of a derivative operator, in the third step we used conditions (b) and (c), and in the fourth step we used the torsion-free condition. Finally, using (d) again, namely $[u, v](f) = [u, v]^a \nabla_a f$, we arrive at (3.1.13). \square

Remark 4 Choose the derivative operator ∂_b of an arbitrary coordinate system $\{x^\mu\}$ as ∇_b from (3.1.13), then we have

$$[u, v]^\mu = (dx^\mu)_a [u, v]^a = u^\nu \partial_\nu v^\mu - v^\nu \partial_\nu u^\mu.$$

This is the claim in Exercise 2.10.

3.2 Derivative and Parallel Transport of a Vector Field Along a Curve

3.2.1 Parallel Transport of a Vector Field Along a Curve

After a derivative operator is assigned to a manifold M , we can introduce the concept of the parallel transport of a vector field along a curve.

Definition 1 Suppose v^a is a vector field along a curve $C(t)$. v^a is said to be **parallelly transported along $C(t)$** if $T^b \nabla_b v^a = 0$, where $T^a \equiv (\partial/\partial t)^a$ is the tangent vector field of the curve.

Just like $T^a \nabla_a f = T(f)$ can be interpreted as the derivative of f along T^a [i.e., along $C(t)$], $T^b \nabla_b v^a$ can be interpreted as the derivative of the vector field v^a along T^a (see

Sect. 3.2.3 for details). Thus, Definition 1 can also be interpreted as: a necessary and sufficient condition for v^a to be parallelly transported along $C(t)$ is that the derivative of it along T^b vanishes.

Theorem 3.2.1 Suppose a curve $C(t)$ is in the coordinate patch of a coordinate system $\{x^\mu\}$ and the parametric representation of the curve is $x^\mu(t)$. Let $T^a \equiv (\partial/\partial t)^a$, then a vector v^a along $C(t)$ satisfies

$$T^b \nabla_b v^a = (\partial/\partial x^\mu)^a (dv^\mu/dt + \Gamma^\mu_{v\sigma} T^v v^\sigma). \quad (3.2.1)$$

Proof Let ∂_a be the ordinary derivative operator of the coordinate system $\{x^\mu\}$, then it follows from (3.1.7) that

$$\begin{aligned} T^b \nabla_b v^a &= T^b (\partial_b v^a + \Gamma^a_{bc} v^c) = T^b [(dx^v)_b (\partial/\partial x^\mu)^a \partial_v v^\mu + \Gamma^a_{bc} v^c] \\ &= T^v (\partial/\partial x^\mu)^a (\partial v^\mu/\partial x^v) + \Gamma^a_{bc} T^b v^c = (\partial/\partial x^\mu)^a [T^v (\partial v^\mu/\partial x^v) + \Gamma^\mu_{v\sigma} T^v v^\sigma], \end{aligned} \quad (3.2.2)$$

where T^v are the coordinate components of the tangent vector T^b of that curve. From (2.2.7) we know that $T^v = dx^v(t)/dt$, and hence

$$T^v (\partial v^\mu/\partial x^v) = [dx^v(t)/dt] [\partial v^\mu(t(x))/\partial x^v] = dv^v(t)/dt.$$

Plugging it into (3.2.2) we obtain (3.2.1). \square

[Optional Reading 3.2.1]

There is one thing that we need to clarify. According to Optional Reading 3.1.1, $\forall p \in C(t)$, to make $\nabla_a v^a|_p$ meaningful v^a needs to be defined at least in a neighborhood U of p . Unfortunately, v^a is only defined on the curve $C(t)$, while any neighborhood of p would contain points that are not on $C(t)$, and thus the $\nabla_b v^a$ in (3.2.1) is actually meaningless! Thankfully, $\nabla_b v^a$ only shows up in the form of $T^b \nabla_b v^a$ in (3.2.1), and $T^b \nabla_b v^a$ does not have this problem. The key point is that adding T^b before $\nabla_b v^a$ tells us to take the derivative of v^a along the tangent direction of the curve, and therefore it only involves the value of v^a on this curve. Now we will explain it precisely. Take the value of (3.2.2) at $p \in C(t)$ we have

$$T^b \nabla_b v^a|_p = (\partial/\partial x^\mu)^a|_p [T^v (\partial v^\mu(x)/\partial x^v) + \Gamma^\mu_{v\sigma} T^v v^\sigma]|_p. \quad (3.2.3)$$

$\partial v^\mu(x)/\partial x^v$ in the square bracket is the derivative of a function v^μ with respect to the argument x^v . When taking the derivative, the tiny change Δx^v of x^v will take the point with coordinates x^v away from p . Since $\Delta x^v (v = 1, \dots, n)$ is arbitrary, the point can move in a neighborhood U of p , which contains points that are not on $C(t)$ whose v^μ are meaningless. Thus, $\partial v^\mu(x)/\partial x^v$ is essentially a meaningless quantity. To resolve this issue, one can define a vector field \bar{v}^a (called the extension of v^a) on the neighborhood U , which is required to be equal to v^a only on $U \cap C(t)$. Now define $T^b \nabla_b \bar{v}^a|_p$ as $T^b \nabla_b \bar{v}^a|_p$, i.e.,

$$T^b \nabla_b v^a|_p \equiv T^b \nabla_b \bar{v}^a|_p = (\partial/\partial x^\mu)^a|_p [T^v (\partial \bar{v}^\mu(x)/\partial x^v) + \Gamma^\mu_{v\sigma} T^v \bar{v}^\sigma]|_p. \quad (3.2.4)$$

Unlike $\partial v^\mu(x)/\partial x^v$, $\partial \bar{v}^\mu(x)/\partial x^v$ has a precise meaning. However, each v^a has an infinite number of extensions \bar{v}^a ; if the $T^b \nabla_b \bar{v}^a|_p$ for different extensions are different, then it would be meaningless to define $T^b \nabla_b v^a|_p$ using (3.2.4). In fact, suppose \bar{v}^a and \tilde{v}^a are two different extensions, then indeed we have $\partial \bar{v}^\mu(x)/\partial x^v \neq \partial \tilde{v}^\mu(x)/\partial x^v$. However, it would no longer

be a problem with T^v added in front of $\partial \bar{v}^\mu(x)/\partial x^v$, since at p we have

$$\begin{aligned} T^v \partial \bar{v}^\mu(x)/\partial x^v &= [\mathrm{d}x^v(t)/\mathrm{d}t][\partial \bar{v}^\mu(x)/\partial x^v] = \mathrm{d}\bar{v}^\mu(t)/\mathrm{d}t \\ &= \mathrm{d}\bar{v}'^\mu(t)/\mathrm{d}t = T^v \partial \bar{v}'^\mu(x)/\partial x^v, \end{aligned}$$

where the key step (the third equality) is because $\bar{v}^\mu(t)$ [the function of one variable that comes from the combination of a vector \bar{v}^a on U and $C(t)$] is equal to $v^\mu(t)$. In conclusion, $\nabla_b v^a|_p$ is meaningless, but $T^b \nabla_b v^a|_p$ is meaningful.

[The End of Optional Reading 3.2.1]

Theorem 3.2.2 *A point $C(t_0)$ on a curve and a vector at this point uniquely defines a vector field that is parallelly transported along the curve.*

Proof If there exists a coordinate system whose coordinate patch contains the whole curve, then it follows from (3.2.1) that $T^b \nabla_b v^a = 0$, the definition of parallel transport, is equivalent to

$$\frac{\mathrm{d}v^\mu}{\mathrm{d}t} + \Gamma^\mu_{\nu\sigma} T^\nu v^\sigma = 0, \quad \mu = 1, \dots, n. \quad (3.2.5)$$

These are n first-order ordinary differential equations of n functions $v^\nu(t)$ to be solved (note that both $\Gamma^\mu_{\nu\sigma}$ and T^ν are given functions of t), and by giving a vector at a point $C(t_0)$ we are actually giving initial conditions $v^\nu(t_0)$ to the equations, and thus there will be a set of unique solutions $v^\nu(t)$. The readers can use the “relay method” to generalize the above proof to the cases where the curve cannot be covered by one coordinate patch. \square

Suppose $p, q \in M$, then V_p and V_q are two vector spaces, and their elements cannot be compared. However, if there is a curve $C(t)$ that connects p and q , we can define a map from V_p to V_q in the following way: $\forall v^a \in V_p$, from Theorem 3.2.2 we know that there is a unique parallelly transported vector field on $C(t)$ (whose value at p is v^a), and its value at q can be defined as the image of v^a . Note that this is a map that depends on the curve, which means v^a could be different for another curve that connects p and q . However, after all, the existence of ∇_a in some ways (although is curve-dependent) connects two vector spaces V_p and V_q ² that were completely unrelated before. Therefore, ∇_a is also called a **connection**.

Beginners often raise questions like: why do we call ∇_a a derivative operator? In other words, why is this ∇_a some kind of generalization of the familiar $\vec{\nabla}$ in 3-dimensional Euclidean space on a general manifold? Why do we interpret $T^b \nabla_b v^a$ as the derivative of v^a along T^b ? Why do we call v^a that satisfies $T^b \nabla_b v^a = 0$ a vector field that is parallelly transported along the curve? In order to answer these questions, we need Sect. 3.2.2 first.

² For the formal definition of a connection given in terms of the language of fiber bundles, see Appendix I in Volume III.

3.2.2 The Derivative Operator Associated with a Metric

Until now, no metric has been involved in this chapter, rather, we only assumed a connection (i.e., a derivative operator) ∇_a is assigned to M . If a metric g_{ab} is also assigned to M , then one can talk about the inner product between two vectors. To make the concept of parallel transport agree with the familiar parallel transport in Euclidean space, we should add the following requirement: suppose u^a and v^a are vector fields parallelly transported along $C(t)$, then $u^a v_a (\equiv g_{ab} u^a v^b)$ is a constant on $C(t)$; that is, the “inner product” of two vectors is invariant under parallel transport. Suppose T^a is the tangent vector field of $C(t)$, then this requirement is equivalent to

$$0 = T^c \nabla_c (g_{ab} u^a v^b) = g_{ab} u^a T^c \nabla_c v^b + g_{ab} v^b T^c \nabla_c u^a + u^a v^b T^c \nabla_c g_{ab} = u^a v^b T^c \nabla_c g_{ab}.$$

A necessary and sufficient condition for the above equation to hold for any curve and any two vector fields that are parallelly transported along the curve is

$$\nabla_c g_{ab} = 0. \quad (3.2.6)$$

When there is no metric, the choice of ∇_c is very arbitrary. After a metric is assigned, we may choose a ∇_c that satisfies the additional requirement $\nabla_c g_{ab} = 0$. Now we will prove that this requirement determines a unique ∇_a .

Theorem 3.2.3 *After assigning a metric g_{ab} to a manifold M , there exists a unique ∇_a such that $\nabla_a g_{bc} = 0$.*

Proof Suppose $\tilde{\nabla}_a$ is an arbitrary derivative operator. We want an appropriate $C^c{}_{ab}$ such that the ∇_a determined by it and $\tilde{\nabla}_a$ satisfies $\nabla_a g_{bc} = 0$. From (3.1.8) we have

$$\nabla_a g_{bc} = \tilde{\nabla}_a g_{bc} - C^d{}_{ab} g_{dc} - C^d{}_{ac} g_{bd} = \tilde{\nabla}_a g_{bc} - C_{cab} - C_{bac}.$$

Hence, it follows from $\nabla_a g_{bc} = 0$ that

$$C_{cab} + C_{bac} = \tilde{\nabla}_a g_{bc}, \quad (3.2.7)$$

Similarly, we have

$$C_{cba} + C_{abc} = \tilde{\nabla}_b g_{ac}, \quad (3.2.8)$$

$$C_{bca} + C_{acb} = \tilde{\nabla}_c g_{ab}. \quad (3.2.9)$$

Adding (3.2.7) to (3.2.8) and subtracting (3.2.9), we obtain by using $C_{cab} = C_{cba}$ that

$$2C_{cab} = \tilde{\nabla}_a g_{bc} + \tilde{\nabla}_b g_{ac} - \tilde{\nabla}_c g_{ab},$$

or

$$C^c{}_{ab} = \frac{1}{2} g^{cd} (\tilde{\nabla}_a g_{bd} + \tilde{\nabla}_b g_{ad} - \tilde{\nabla}_d g_{ab}). \quad (3.2.10)$$

The combination of this $C^c{}_{ab}$ and $\tilde{\nabla}_a$, namely ∇_a , is then the solution to the equation $\nabla_a g_{bc} = 0$. This must be the unique solution since if ∇'_a also satisfies $\nabla'_a g_{bc} = 0$, treating ∇'_a as $\tilde{\nabla}_a$ we can see that $C^c{}_{ab}$ vanishes, which means there is no difference between ∇_a and ∇'_a . \square

The ∇_a that satisfies $\nabla_a g_{bc} = 0$ is called **the derivative operator associated (or compatible) with g_{bc}** . From now on, unless stated otherwise, when we talk about ∇_a when there is a g_{ab} , we will choose it to be the derivative operator associated with g_{ab} . It can be proved that (exercise) $\nabla_a g_{bc} = 0$ assures that $\nabla_a g^{bc} = 0$ (and vice versa), which is exceptionally convenient for performing calculations.

Example 1 In Euclidean space, there exist infinitely many derivative operators that satisfy Definition 1 in Sect. 3.1. However, there is only one derivative operator that is associated with the Euclidean metric δ_{ab} , namely the ordinary derivative operator ∂_a of a Cartesian coordinate system $\{x^\mu\}$ (all Cartesian systems have the same ∂_a), since it follows from the definition of δ_{ab} (2.5.11) that $\partial_c \delta_{ab} = (dx^\sigma)_c (dx^\mu)_a (dx^\nu)_b \partial_\sigma \delta_{\mu\nu} = 0$. For the 3-dimensional Euclidean space, the ∂_a of a Cartesian coordinate system is the familiar $\vec{\nabla}$ in the standard vector field theory.

Suppose ∇_a is associated with g_{bc} , and choose the ∂_a of an arbitrary coordinate system as $\tilde{\nabla}_a$, then the $C^c{}_{ab}$ in (3.2.10) is the Christoffel symbol $\Gamma^c{}_{ab}$ of ∇_a in this coordinate system. From this equation it is not difficult to derive the following expression for the components $\Gamma^\sigma{}_{\mu\nu}$ of $\Gamma^c{}_{ab}$ in this system:

$$\Gamma^\sigma{}_{\mu\nu} = \frac{1}{2} g^{\sigma\rho} (g_{\rho\mu,\nu} + g_{\nu\rho,\mu} - g_{\mu\nu,\rho}). \quad (3.2.10')$$

The derivation is as follows: regard ∂_a as the $\tilde{\nabla}_a$ in (3.2.10), and then the $C^c{}_{ab}$ in the equation is $\Gamma^c{}_{ab}$; hence,

$$\begin{aligned} \Gamma^c{}_{ab} &= \frac{1}{2} g^{cd} (\partial_a g_{bd} + \partial_b g_{ad} - \partial_d g_{ab}), \\ \Gamma^\sigma{}_{\mu\nu} &= \Gamma^c{}_{ab} (dx^\sigma)_c (\partial/\partial x^\mu)^a (\partial/\partial x^\nu)^b \\ &= \frac{1}{2} (dx^\sigma)_c (\partial/\partial x^\mu)^a (\partial/\partial x^\nu)^b g^{cd} (\partial_a g_{bd} + \partial_b g_{ad} - \partial_d g_{ab}) \\ &= \frac{1}{2} g^{\sigma\rho} (\partial_\mu g_{\nu\rho} + \partial_\nu g_{\mu\rho} - \partial_\rho g_{\mu\nu}) = \frac{1}{2} g^{\sigma\rho} (g_{\rho\mu,\nu} + g_{\nu\rho,\mu} - g_{\mu\nu,\rho}). \end{aligned}$$

[We used $\partial_a (\partial/\partial x^\nu)^b = 0$ in the second-to-last equality.] Using the symmetry $\Gamma^\sigma{}_{\mu\nu} = \Gamma^\sigma{}_{\nu\mu}$ it is not difficult to see that this is exactly (3.2.10'). By combining this equation and (3.1.11), it is easy to derive the coordinate components $v^\nu{}_{;\mu}$ and $\omega_{\nu;\mu}$ of covariant derivatives $\nabla_a v^b$ and $\nabla_a \omega_b$.

Remark 1 $\Gamma^\sigma_{\mu\nu}$ depends on both the ∇_a assigned on M and the coordinate system. If M has a metric g_{ab} , then ∇_a will be referred to as the derivative operator associated with g_{ab} except when otherwise stated, and “the Christoffel symbol of the coordinate system” will refer to the Christoffel symbol of this ∇_a in this coordinate system. For instance, when talking about the Christoffel symbol of a coordinate system in the 3-dimensional Euclidean space, we mean the Christoffel symbol of the ∇_a associated with the Euclidean metric (i.e., the ∂_a of a Cartesian system) in this coordinate system. The Christoffel symbol of the ∂_a of a Cartesian system is obviously zero in any Cartesian system. As an exercise (Exercise 3.7), the reader may derive all the nonvanishing $\Gamma^\sigma_{\mu\nu}$ of ∂_a in the spherical coordinate system using (3.2.10').

Suppose \vec{T} is a vector field in the 3-dimensional Euclidean space, then $\vec{T} \cdot \vec{\nabla} f$ is the component of $\vec{\nabla} f$ in the direction of \vec{T} , i.e., the derivative of f along \vec{T} . On the other hand, it follows from condition (d) of a derivative operator that $T^a \partial_a f = T(f)$, and the right-hand side of it is exactly the derivative of f along T^a . Thus, $T^a \partial_a f = \vec{T} \cdot \vec{\nabla} f$. A further question is: what does $T^b \partial_b v^a$ stand for? The answer is that it stands for the derivative of v^a along T^a . For the details, see Sect. 3.2.3.

3.2.3 Relationship Between the Derivative and Parallel Transport of a Vector Field Along a Curve

First we talk about the simplest case, i.e., Euclidean space. There is one type of special coordinate system (Cartesian system) in Euclidean space, using which we can define the absolute (curve-independent) parallel transport of a vector.

Definition 2 A vector $\vec{\tilde{v}}$ at p in Euclidean space is referred to as the result of a vector \vec{v} at q parallelly transported to p if their components in the same Cartesian system are the same. (NB: The parallel transport for one Cartesian system is the parallel transport for all Cartesian systems.)

Definition 3 In Euclidean space, the derivative of a vector field \vec{v} on a curve $C(t)$ along the curve, denoted by $d\vec{v}/dt$, is defined as

$$\frac{d\vec{v}}{dt} \Big|_p := \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} (\vec{\tilde{v}}|_p - \vec{v}|_p) \quad \forall p \in C(t), \quad (3.2.11)$$

where $\vec{\tilde{v}}|_p$ is the result of $\vec{v}|_q$ parallelly transported to p (q is a neighboring point of p on the curve), and $\Delta t \equiv t(q) - t(p)$. Now we will show that $d\vec{v}/dt$ is $T^b \partial_b v^a$ in the abstract index notation [where T^b is the tangent vector field of $C(t)$, and ∂_b is the ordinary derivative operator of a Cartesian system], to do which we only have to show that their components are the same in a Cartesian system $\{x^i\}$:

$$\text{the } i\text{th component of } T^b \partial_b v^a = (dx^i)_a T^b \partial_b v^a = T^b \partial_b [(dx^i)_a v^a] = T^b \partial_b v^i = T(v^i) = \frac{dv^i}{dt}.$$

[where we used condition (d) of a derivative operator in the fourth equality, and the definition of a tangent vector, (2.2.6'), in the fifth equality.] On the other hand, from (3.2.11) we can see that

$$\text{the } i\text{th component of } \frac{d\vec{v}}{dt} \Big|_p = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} (\tilde{v}^i|_p - v^i|_p) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} (v^i|_q - v^i|_p) = \frac{dv^i}{dt} \Big|_p ,$$

[where we used Definition 2 in the second equality, and the third equality is nothing but the definition of the derivative of a function $v^i(t)$.] Comparing the two equations we can see that

$$\frac{d\vec{v}}{dt} = T^b \partial_b v^a . \quad (3.2.12)$$

Generalizing to any curve $C(t)$ on any manifold M with any ∇_a , we can naturally call $T^b \nabla_b v^a$ the derivative of v^a along T^b [or along $C(t)$]. Sometimes this derivative is also denoted by Dv^a/dt , i.e.,

$$\frac{Dv^a}{dt} \equiv T^b \nabla_b v^a . \quad (3.2.13)$$

However, Definition 2 cannot be generalized to an arbitrary manifold (M, ∇_a) with an arbitrary connection. In Sect. 3.4 we will introduce the concept of the intrinsic curvature of a manifold, and it will be pointed out in Sect. 3.5 that only a space whose intrinsic curvature is zero has the concept of absolute (i.e., curve-independent) parallel transport. However, seeing that $d\vec{v}/dt$ is equivalent to \vec{v} parallelly transported along $C(t)$ in Euclidean space, we can naturally regard $T^b \nabla_b v^a = 0$ as the definition of the parallel transport of a vector field v^a on a curve $C(t)$ in (M, ∇_a) . This explains the motivation of Definition 1 in Sect. 3.2.1 (although one should pay attention that the parallel transport defined in this way usually depends on the curve). In this manner, we first defined $T^b \nabla_b v^a$, the derivative of v^a along the curve, then using which defined the parallel transport of v^a along the curve (which is in the opposite order to how we treat a Euclidean space). Since the derivative $T^b \nabla_b v^a$ of v^a along the curve is kind of abstract, after having the curve-dependent notion of parallel transport, it is beneficial to interpret $T^b \nabla_b v^a$ by aid of the term parallel transport. The essence of this interpretation is actually the meaning of (3.2.11), see the following theorem.

Theorem 3.2.4 Suppose v^a is a vector field on a curve $C(t)$ of (M, ∇_a) , T^b is the tangent vector field of $C(t)$, p and q are neighboring points on $C(t)$ (see Fig. 3.1), then

$$T^b \nabla_b v^a|_p = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} (\tilde{v}^a|_p - v^a|_p) , \quad (3.2.14)$$

where $\Delta t \equiv t(q) - t(p)$, and $\tilde{v}^a|_p$ is the result of $v^a|_q$ parallelly transported to p .

Fig. 3.1 Parallelly transporting $v^a|_q$ along the curve to p yields $\tilde{v}^a|_p$, we can subtract $v^a|_p$ from it and define the derivative along the curve

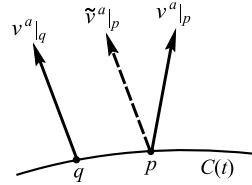
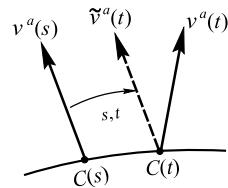


Fig. 3.2 $\psi_{s,t}$ maps $v^a(s)$ to $\tilde{v}^a(t)$



Proof [Optional Reading]

All we have to prove is the following equivalent statement:

$$T^b \nabla_b v^a|_t = \frac{d}{ds} [\psi_{s,t} v(s)]^a \Big|_{s=t}, \quad (3.2.15)$$

where $T^b \nabla_b v^a|_t$ and $v^a(s)$ are short for $T^b \nabla_b v^a|_{C(t)}$ and $v^a(C(s))$, respectively, and $\psi_{s,t}$ is the translation map from vector space $V_{C(s)}$ to $V_{C(t)}$ (see Fig. 3.2). It is not difficult to show that (Exercise 3.8) $\psi_{s,t} : V_{C(s)} \rightarrow V_{C(t)}$ is an isomorphism.

Suppose \tilde{v}^a is a vector field parallelly transported along $C(t)$ that is determined by $v^a(s)$, then

$$\tilde{v}^a(t) = [\psi_{s,t} v(s)]^a, \quad (3.2.16)$$

$$T^b \nabla_b \tilde{v}^a = 0. \quad (3.2.17)$$

The coordinate component expression for (3.2.16) is

$$\tilde{v}^\mu(t) = (\psi_{s,t})^\mu_\nu v^\nu(s), \quad (3.2.16')$$

where $(\psi_{s,t})^\mu_\nu$ are the elements of the matrix $(\psi_{s,t})$. The coordinate component expression for (3.2.17) is

$$\frac{d\tilde{v}^\mu(t)}{dt} + \Gamma^\mu_{\nu\sigma} T^\sigma \tilde{v}^\nu = 0.$$

Using (3.2.16') this equation can also be written as

$$\frac{d}{dt} [(\psi_{s,t})^\mu_\nu v^\nu(s)] + \Gamma^\mu_{\nu\sigma} T^\sigma (\psi_{s,t})^\nu_\rho v^\rho(s) = 0.$$

Applying the above equation to $t = s$, and noticing that $\psi_{s,s}$ is the identity map, we obtain

$$\frac{d}{dt} [(\psi_{s,t})^\mu_\nu] \Big|_{t=s} = -(\Gamma^\mu_{\nu\sigma} T^\sigma)|_s. \quad (3.2.18)$$

On the other hand, by definition, $\psi_{t,s}$ is the inverse map of $\psi_{s,t}$, i.e., $(\psi_{s,t})^\mu_\rho (\psi_{t,s})^\rho_\nu = \delta^\mu_\nu$, and hence

$$\begin{aligned} 0 &= \left[\frac{d(\psi_{s,t})^\mu}{ds} \rho (\psi_{t,s})^\rho v \right] \Big|_{s=t} + \left[(\psi_{s,t})^\mu \rho \frac{d(\psi_{t,s})^\rho}{ds} v \right] \Big|_{s=t} \\ &= \frac{d(\psi_{s,t})^\mu}{ds} v \Big|_{s=t} + \frac{d(\psi_{t,s})^\mu}{ds} v \Big|_{s=t}. \end{aligned} \quad (3.2.19)$$

Now let us prove (3.2.15). The μ th component of the right-hand side of this equation is

$$\begin{aligned} \frac{d}{ds} [\psi_{s,t} v(s)]^\mu \Big|_{s=t} &= \frac{d}{ds} [(\psi_{s,t})^\mu v^\nu(s)] \Big|_{s=t} \\ &= \frac{d}{ds} (\psi_{s,t})^\mu v \Big|_{s=t} + (\psi_{s,t})^\mu v \Big|_{s=t} \frac{dv^\nu(s)}{ds} \Big|_{s=t} \\ &= -\frac{d}{ds} (\psi_{t,s})^\mu v \Big|_{s=t} + \delta^\mu_\nu \frac{dv^\nu(s)}{ds} \Big|_{s=t} \\ &= (\Gamma^\mu_{\nu\sigma} T^\sigma) |_t v^\nu(t) + \frac{dv^\mu(s)}{ds} \Big|_{s=t} \\ &= (\Gamma^\mu_{\nu\sigma} T^\sigma v^\nu) |_t + \frac{dv^\mu(s)}{ds} \Big|_{s=t} = (T^b \nabla_b v^\mu) |_t, \end{aligned}$$

where we used (3.2.19) in the third step and (3.2.18) in the fourth step. The right-hand side of the above equation is the μ th component of the right-hand side of (3.2.15), and (3.2.14) is therefore proved. \square

3.3 Geodesics

Definition 1 A curve $\gamma(t)$ on (M, ∇_a) is called a **geodesic** if its tangent vector field T^a satisfies $T^b \nabla_b T^a = 0$.

Remark 1 ① We can see that a necessary and sufficient condition for a curve to be a geodesic is that its tangent vector field is parallelly transported along the curve. ② $T^b \nabla_b T^a = 0$ is called a **geodesic equation**. ③ Suppose there is a metric field g_{ab} on a manifold M , then the geodesics of (M, g_{ab}) refer to the geodesics of (M, ∇_a) , where ∇_a is associated with g_{ab} .

Suppose a geodesic $\gamma(t)$ is located in the coordinate patch of a coordinate system, then substituting T^a for the v^a in (3.2.5) yields

$$\frac{dT^\mu}{dt} + \Gamma^\mu_{\nu\sigma} T^\nu T^\sigma = 0, \quad \mu = 1, \dots, n.$$

Suppose $x^\nu = x^\nu(t)$ are the parametric equations of $\gamma(t)$, then $T^\mu = dx^\mu/dt$. Hence, the equation above can be rewritten as

$$\frac{d^2 x^\mu}{dt^2} + \Gamma^\mu_{\nu\sigma} \frac{dx^\nu}{dt} \frac{dx^\sigma}{dt} = 0, \quad \mu = 1, \dots, n. \quad (3.3.1)$$

This is the coordinate component expression for a geodesic equation.

Example 1 The Christoffel symbol of the Euclidean (Minkowski) metric in a Cartesian (Lorentzian) system vanishes, and the general solution to the geodesic equation (3.3.1) is $x^\mu(t) = a^\mu t + b^\mu$ (where a^μ, b^μ are constants). If we call the curve in Euclidean (Minkowski) space with parametric equations $x^\mu(t) = a^\mu t + b^\mu$ a straight line (segment), then a geodesic in Euclidean (Minkowski) space is synonymous with a straight line (segment). Thus, a geodesic can be viewed as the generalization of the concept of a straight line in Euclidean space to a generalized Riemannian space.

Example 2 Suppose S^2 is a 2-dimensional sphere in a 3-dimensional Euclidean space. Set up a spherical coordinate system that is centered at the origin, then the 3-dimensional Euclidean line element is $ds^2 = dr^2 + r^2(d\theta^2 + \sin^2 \theta d\phi^2)$. If the line element is lying on S^2 then $r = R$ (the radius of the sphere) leads to $dr = 0$, and hence the “induced line element” (called the standard spherical line element) is $d\hat{s}^2 = R^2(d\theta^2 + \sin^2 \theta d\phi^2)$. That is, the 3-dimensional Euclidean metric δ_{ab} induces a 2-dimensional metric g_{ab} on a sphere S^2 , whose components in the coordinate basis $\{(\partial/\partial\theta)^a, (\partial/\partial\phi)^a\}$ are $g_{\theta\theta} = R^2$, $g_{\phi\phi} = R^2 \sin^2 \theta$, $g_{\theta\phi} = g_{\phi\theta} = 0$. It can be proved from (3.1.1) that, when measured by this metric, a curve on the sphere is a geodesic if and only if it is a great circle (with an appropriate parametrization).

Theorem 3.3.1 Suppose $\gamma(t)$ is a geodesic, then the tangent vector field T'^a of its reparametrization $\gamma'(t')$ [$= \gamma(t)$] satisfies

$$T'^b \nabla_b T'^a = \alpha T'^a \quad [\alpha \text{ is a function defined on } \gamma(t)]. \quad (3.3.2)$$

Proof

$$\begin{aligned} T^a &= \left(\frac{\partial}{\partial t} \right)^a = \left(\frac{\partial}{\partial t'} \right)^a \frac{dt'}{dt} = \frac{dt'}{dt} T'^a, \\ 0 &= T^b \nabla_b T^a = \frac{dt'}{dt} T'^b \nabla_b \left(\frac{dt'}{dt} T'^a \right) = \left(\frac{dt'}{dt} \right)^2 T'^b \nabla_b T'^a + T'^a \frac{dt'}{dt} T'^b \nabla_b \left(\frac{dt'}{dt} \right) \\ &= \left(\frac{dt'}{dt} \right)^2 T'^b \nabla_b T'^a + T'^a \frac{dt'}{dt} \frac{d}{dt'} \left(\frac{dt'}{dt} \right) = \left(\frac{dt'}{dt} \right)^2 T'^b \nabla_b T'^a + T'^a \frac{d^2 t'}{dt^2}, \end{aligned}$$

and hence $T'^b \nabla_b T'^a = - \left(\frac{dt'}{dt} \right)^2 \frac{d^2 t'}{dt^2} T'^a$. Set $\alpha \equiv - \left(\frac{dt'}{dt} \right)^2 \frac{d^2 t'}{dt^2}$, then (3.3.2) is satisfied. \square

Theorem 3.3.2 Suppose the tangent vector field T^a of a curve $\gamma(t)$ satisfies $T^b \nabla_b T^a = \alpha T^a$ [α is a function on $\gamma(t)$], then there exists a $t' = t'(t)$ such that $\gamma'(t') [= \gamma(t)]$ is a geodesic.

Proof Exercise 3.9. \square

Definition 2 A parameter which makes a curve become a geodesic is called an **affine parameter** of this curve.

Remark 2 Sometimes a curve that satisfies $T^b \nabla_b T^a = \alpha T^a$ is also called a geodesic. Nonetheless, in order to avoid confusion, a better way to call it is a “non-affinely parametrized geodesic”.

Theorem 3.3.3 *If t is an affine parameter of a geodesic, then the necessary and sufficient condition of any parameter t' of this curve to be an affine parameter is $t' = at + b$ (where a, b are constants and $a \neq 0$).*

Proof Exercise 3.9. □

Theorem 3.3.4 *A point p of a manifold (M, ∇_a) with connection and a vector v^a at p determines the unique geodesic $\gamma(t)$ that satisfies*

- (1) $\gamma(0) = p$;
- (2) *the tangent vector of $\gamma(t)$ at p is equal to v^a .*

Proof Choose an arbitrary coordinate system $\{x^\mu\}$ whose coordinate patch contains p , then the definition of a geodesic is equivalent to (3.3.1). Consider this equation as n second-order ordinary differential equations with respect to n unknown functions $x^\mu(t)$, then giving $p \in M$ and $v^a \in V_p$ is giving the initial conditions $x^\mu(0) = x^\mu|_p$ and $(dx^\mu/dt)|_0 = v^\mu$, and hence there is a unique solution. □

Remark 3 Similar to Theorem 2.2.8, the word “unique” in Theorem 3.3.4 should also be understood as “locally unique”.

The discussions above do not involve a metric. From now on, we will suppose there is a metric field g_{ab} on M . Since a tangent vector T^a is parallelly transported along a geodesic, and since the self “inner product” $g_{ab}T^aT^b$ of a parallelly transported vector is a constant, the sign of $g_{ab}T^aT^b$ does not change along the geodesic, which indicates that geodesics can always be classified as three types: timelike, spacelike and null (there is no “outlandish” geodesic that can turn from one type into another).

Theorem 3.3.5 *The arc length parameter of a (nonnull) geodesic is an affine parameter.*

Proof Exercise 3.9. Hint: First show that a tangent vector of an affinely parametrized geodesic has a constant magnitude along the curve. □

As we all know, a straight line (segment) is the shortest path between two points in Euclidean space. Now we will discuss to what extent this conclusion can be applied to a manifold with a Lorentzian metric (a spacetime).

Theorem 3.3.6 *Suppose g_{ab} is a Lorentzian metric field on a manifold M , and $p, q \in M$, then a smooth spacelike (timelike) curve between p and q is a geodesic if and only if it extremizes the arc length of the curve.*

Remark 4 ① This theorem also holds for any case where g_{ab} is positive definite [in this case the modifier “spacelike (timelike)” is omitted]. ② The meaning of extremizing the arc length is as follows: suppose C is a spacelike (timelike) curve between

p and q , then one can add a small modification to it and obtain many spacelike (timelike) curves that are “infinitely close” to C . Theorem 3.3.6 claims that, a necessary and sufficient condition for a curve C to be a geodesic is that the length of the curve is an extremum among the lengths of all possible spacelike (timelike) curves. The condition for a function $f(x)$ of one variable to take an extremum is that its first order derivative is zero. However, the “argument” corresponds to the length l (which can be seen as the “function value”) in Theorem 3.3.6 is not a real number but a curve. Here we are concerned about the change of l when a curve turns into another curve, and thus l is not a function but a functional. According to the theory of variations, the necessary and sufficient condition for l to be extremized is that its variation δl vanishes.

Proof [Optional Reading]

We will give a proof by means of a coordinate system. Suppose $C(t)$ is a curve, $x^\mu(t)$ are the parametric equations of it in a coordinate system, $p \equiv C(t_1)$ and $q \equiv C(t_2)$, then it follows from (2.5.6) that the arc length from p to q can be expressed using the coordinate language as

$$l = \int_{t_1}^{t_2} \left(g_{\mu\nu} \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} \right)^{1/2} dt. \quad (3.3.3)$$

[We assume that $C(t)$ is a spacelike curve. If $C(t)$ is timelike, then a minus sign should be added in the parentheses of this equation, which does not affect the result.] Suppose $C'(t)$ is an “infinitely close” spacelike curve whose parametrization $x'^\mu(t)$ satisfy $x'^\mu(t_1) = x^\mu(t_1)$, $x'^\mu(t_2) = x^\mu(t_2)$, and the variation $\delta x^\mu(t) \equiv x'^\mu(t) - x^\mu(t)$ is “infinitesimal”. This variation causes a small change in $g_{\mu\nu}$ and the components of the tangent dx^μ/dt as

$$\delta g_{\mu\nu} \equiv g_{\mu\nu}[x^\sigma(t) + \delta x^\sigma(t)] - g_{\mu\nu}[x^\sigma(t)] = \frac{\partial g_{\mu\nu}}{\partial x^\sigma} \delta x^\sigma(t)$$

and

$$\delta \left(\frac{dx^\mu}{dt} \right) \equiv \frac{d(x^\mu + \delta x^\mu)}{dt} - \frac{dx^\mu}{dt} = \frac{d(\delta x^\mu)}{dt},$$

which, through (3.3.3), give rise to the following variation of l :

$$\begin{aligned} \delta l &= \frac{1}{2} \int_{t_1}^{t_2} \left(g_{\mu\nu} \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} \right)^{-1/2} \\ &\quad \times \left[g_{\mu\nu} \frac{dx^\mu}{dt} \frac{d}{dt}(\delta x^\nu) + g_{\mu\nu} \frac{dx^\nu}{dt} \frac{d}{dt}(\delta x^\mu) + \frac{\partial g_{\mu\nu}}{\partial x^\sigma} (\delta x^\sigma) \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} \right] dt. \end{aligned}$$

Since arc length is independent of the parametrization of a curve, one can choose the most convenient parameter for the calculation. Theorem 3.3.5 indicates that no matter what the old parameter is (denoted by \tilde{t} for now), we can always choose a new parameter $t = t(\tilde{t})$ such that the length of the tangent vector at each point of the curve is normalized, i.e., $g_{\mu\nu} \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} = 1$ (namely the arc length parameter). Also, noticing the symmetry of $g_{\mu\nu}$, the equation above can then be simplified as

$$\begin{aligned}\delta l &= \int_{t_1}^{t_2} \left[g_{\mu\nu} \frac{dx^\mu}{dt} \frac{d}{dt} (\delta x^\nu) + \frac{1}{2} \frac{\partial g_{\mu\nu}}{\partial x^\sigma} (\delta x^\sigma) \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} \right] dt \\ &= \int_{t_1}^{t_2} \left[\frac{d}{dt} \left(g_{\mu\nu} \frac{dx^\mu}{dt} \delta x^\nu \right) - \frac{d}{dt} \left(g_{\mu\nu} \frac{dx^\mu}{dt} \right) \delta x^\nu + \frac{1}{2} \frac{\partial g_{\mu\nu}}{\partial x^\sigma} (\delta x^\sigma) \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} \right] dt \\ &= \int_{t_1}^{t_2} \left[-\frac{d}{dt} \left(g_{\mu\nu} \frac{dx^\mu}{dt} \right) + \frac{1}{2} \frac{\partial g_{\mu\nu}}{\partial x^\sigma} \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} \right] (\delta x^\sigma) dt,\end{aligned}$$

where in the last step we used the premise that δx^σ vanishes at $C(t_1)$ and $C(t_2)$. The equation above indicates that the necessary and sufficient condition for δl to vanish for any δx^σ is that

$$\begin{aligned}0 &= -\frac{d}{dt} \left(g_{\mu\nu} \frac{dx^\mu}{dt} \right) + \frac{1}{2} \frac{\partial g_{\mu\nu}}{\partial x^\sigma} \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} \\ &= -g_{\mu\nu} \frac{d^2 x^\mu}{dt^2} - \frac{\partial g_{\mu\nu}}{\partial x^\nu} \frac{dx^\nu}{dt} \frac{dx^\mu}{dt} + \frac{1}{2} \frac{\partial g_{\mu\nu}}{\partial x^\sigma} \frac{dx^\mu}{dt} \frac{dx^\nu}{dt}.\end{aligned}$$

Contracting this equation with $g^{\rho\sigma}$ yields

$$\begin{aligned}0 &= -\frac{d^2 x^\rho}{dt^2} - g^{\rho\sigma} (g_{\mu\sigma,\nu} - \frac{1}{2} g_{\mu\nu,\sigma}) \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} \\ &= -\frac{d^2 x^\rho}{dt^2} - \frac{1}{2} g^{\rho\sigma} (g_{\sigma\mu,\nu} + g_{\nu\sigma,\mu} - g_{\mu\nu,\sigma}) \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} \\ &= -\frac{d^2 x^\rho}{dt^2} - \Gamma^\rho_{\mu\nu} \frac{dx^\mu}{dt} \frac{dx^\nu}{dt}.\end{aligned}$$

This is exactly the coordinate expression for the geodesic equation (3.3.1).

As a problem for thinking, the reader may consider what result it leads to if we do not set $g_{\mu\nu} \frac{dx^\mu}{dt} \frac{dx^\nu}{dt} = 1$. \square

An extremum of a function of one variable can be either a minimum [sufficient condition is $f'(x) = 0, f''(x) > 0$], a maximum [sufficient condition is $f'(x) = 0, f''(x) < 0$] or neither of these [necessary condition is $f'(x) = 0, f''(x) = 0$]. Similar to this, the extremum of arc length also has the above three possibilities, which we shall discuss below.

First, we discuss the case where g_{ab} is positive definite. Given an arbitrary curve between p and q , one can always modify it a little and obtain a curve with greater length, and hence there is no maximum length of a curve between p and q . Suppose C is a curve between p and q with minimum length, then it follows Theorem 3.3.6 that it must be a geodesic. However, the length of a geodesic between p and q is not necessarily minimum since an extremum can be neither a minimum nor a maximum. For instance, Fig. 3.3 represents a sphere, γ_1 and γ_2 are two geodesics from the south pole s to the north pole n that are very close to each other, and γ is another geodesic. Although the curve $sand$ is a geodesic between s and d , its length is not a minimum. The point is that there is a north pole n on the curve, which is “conjugate” to the south pole s ; that is, there exists a geodesic γ_2 from s to n that is “infinitely close” to γ_1 (for the precise definition of a pair of “conjugate points”, see Optional Reading 7.6.3). It can be proved that the necessary and sufficient condition for the length of a geodesic to be minimum is that there is no pair of conjugate points on

Fig. 3.3 The length of a geodesic $s\gamma d$ is not a minimum

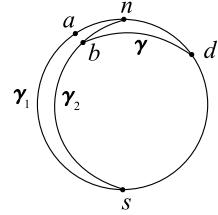
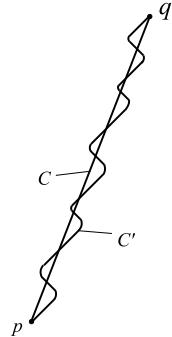


Fig. 3.4 Given a timelike curve C between p and q , one can always find a nearby timelike curve C' shorter than it

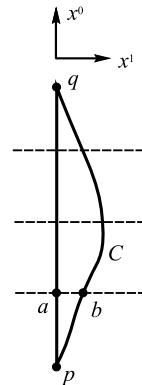


the curve. There is certainly no conjugate points in Euclidean space, and therefore a straight line (segment) is the shortest between two points.

And then we discuss the case where g_{ab} is a Lorentzian metric. We first look at Minkowski spacetime as the simplest example. We have said that straight lines and geodesics are synonymous in Minkowski spacetime. Suppose p and q are connected by a timelike geodesic γ . Is it the shortest curve between p and q ? No. Since the length of a null curve is zero, any timelike curve C is not the shortest. One can always modify it slightly and make it a timelike curve C' that is close enough to null whose length is less than C (see Fig. 3.4). In fact, not only is a timelike geodesic γ not the shortest, but it is also the longest curve between p and q . Here we show it in 2-dimensional Minkowski spacetime as an example (it can be easily carried over to an arbitrary dimensional Minkowski spacetime). Since the parametric representation $x^\mu(t)$ of γ are linear functions, by performing a translation and a boost [(2.5.19), (2.5.20)] of the Lorentzian coordinates, we can choose a Lorentzian system $\{x^0, x^1\}$ that can make the coordinate line of x^0 coincide with γ . Suppose C is an arbitrary timelike non-geodesic between p and q , we can use a lot of constant- x^0 lines to divide γ into many line segments (see Fig. 3.5). From the expression for a Minkowski line element we can see that the arc length of the line segments pa and pb , respectively, are

$$\begin{aligned} dl_{pa} &= \sqrt{-ds^2} = \sqrt{-[-(dx^0)^2 + 0]} = dx^0, \\ dl_{pb} &= \sqrt{-[-(dx^0)^2 + (dx^1)^2]} < dx^0 = dl_{pa}. \end{aligned}$$

Fig. 3.5 A geodesic γ is the longest timelike curve between p and q



This result can also be applied to any other line segment, and thus $l_\gamma > l_C$, i.e., a timelike geodesic is the longest timelike curve between two points in Minkowski spacetime. In other words, a (timelike) straight line (segment) is the longest between two points in Minkowski spacetime. And since the longest curve must be a geodesic, the necessary and sufficient condition for a timelike curve between two points in a Minkowski spacetime to be the longest is that it is a geodesic. Now let us talk about a general spacetime. Suppose C is the timelike curve between p and q that has the greatest length, then it follows from Theorem 3.3.6 that it is a geodesic. However, the converse is not necessarily true, because Theorem 3.3.6 only assures that the length of a geodesic between p and q is an extremum, but does not guarantee that it is a maximum. (Of course, it is definitely not a minimum either since the length of a null curve is zero.) It can be proved that the necessary and sufficient condition for the length of a geodesic in an arbitrary spacetime to be a maximum is that there is no pair of conjugate points on the curve. Summary: for two points that are timelike related in any spacetime: ① the longest curve between them is a timelike geodesic; ② a timelike geodesic between them is not necessarily the longest curve (though for Minkowski spacetime it certainly is); ③ there is no shortest timelike curve between them.

[Optional Reading 3.3.1]

Using geodesics we can define two useful concepts; namely, the exponential map of a generalized Riemannian space (M, g_{ab}) and Riemannian normal coordinates.

The **exponential map** of $p \in M$ is a map from V_p (or a subset of it) to a manifold M , denoted by

$$\exp_p : V_p \text{ (or a subset of it)} \rightarrow M,$$

defined as follows: $\forall v^a \in V_p$, (p, v^a) determines a unique geodesic $\gamma(t)$. If we set the affine parameter t as zero at p , then the image of v^a under the map \exp_p is defined as the point with $t = 1$ on the geodesic, i.e., $\exp_p(v^a) := \gamma(1)$. Suppose $\underline{0}$ is the zero element of V_p . Since the unique geodesic determined by $(p, 0)$ maps all the points of \mathbb{R} (or an interval of it) to p , we have $\exp_p(\underline{0}) = p$. However, if we remove the point $\gamma(1)$ from M , i.e., we use $M - \{\gamma(1)\}$ as the background manifold (see Fig. 3.6), then v^a has no image under the map \exp_p . Therefore, the domain of the exponential map can only be a subset of V_p , denoted by

Fig. 3.6 Removing a point $\gamma(1)$, then v^a has no image under \exp_p

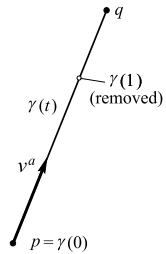


Fig. 3.7 Two geodesics determined by (p, v^a) and (p, v'^a) intersect at q . Choosing the magnitude of v^a and v'^a such that $q = \gamma(1) = \gamma'(1)$, we can see that \exp_p is not one-to-one

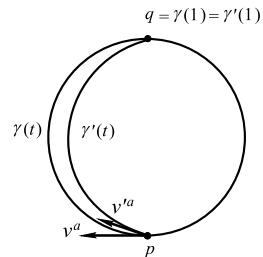
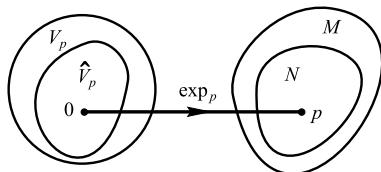


Fig. 3.8 $\exp_p : \hat{V}_p \rightarrow N$ is a diffeomorphism



\hat{V}_p , i.e., $\exp_p : \hat{V}_p \rightarrow M$. Figure 3.7 indicates that two geodesics $\gamma(t)$ and $\gamma'(t)$ determined by (p, v^a) and (p, v'^a) intersect at q . Choosing the magnitude of v^a and v'^a appropriately, one can make $q = \gamma(1) = \gamma'(1)$, so that

$$q = \exp_p(v^a) = \exp_p(v'^a).$$

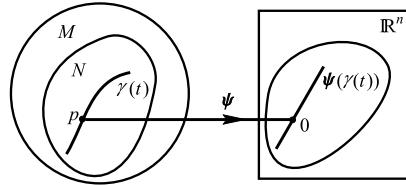
Thus, in this case \exp_p is not a one-to-one map. Since we have removed a point as shown in Fig. 3.6, there is no $u^a \in V_p$ for q such that $q = \exp_p(u^a)$; thus, in this case \exp_p is not an onto map. However, it can be proved that as long as we add proper constraints on the domain and the range of \exp_p , it will be not only one-to-one and onto, but also a diffeomorphism. See the following theorem:

Theorem 3.3.7 *∀p ∈ M, one can always find an open subset \hat{V}_p that contains the zero element in the tangent space V_p of p (regarded as an n dimensional manifold), and find an open subset N of M that contains p such that $\exp_p : \hat{V}_p \rightarrow N$ is a diffeomorphism (see Fig. 3.8).*

Proof See Hawking and Ellis (1973) pp. 33–34. □

Definition 3 A neighborhood N of $p \in M$ is called a **normal neighborhood** of p if V_p has an open subset \hat{V}_p such that $\exp_p : \hat{V}_p \rightarrow N$ is a diffeomorphism.

Fig. 3.9 Figure of Theorem 3.3.8



Using this diffeomorphism $\exp_p : \hat{V}_p \rightarrow N$ we can define coordinates inside N : choose an arbitrary basis $\{(e_\mu)^a\}$ of V_p , and define the n components of $v^a \equiv \exp_p^{-1}(q) \in \hat{V}_p$, the inverse image of $q \in N$ under \exp_p , as the n coordinates of q . The coordinate system defined in this way is called the **Riemannian normal coordinate system**, whose coordinate patch is N .

Theorem 3.3.8 Suppose (N, ψ) is a Riemannian normal coordinate system of a point p , then the image of each geodesic $\gamma(t)$ in N that passes through p under the map ψ , denoted by $\psi(\gamma(t))$, is a straight line in \mathbb{R}^n that passes through the origin (see Fig. 3.9).

Proof Without loss of generality, we may consider $p = \gamma(0)$. Denote $v_1^a \equiv (\partial/\partial t)^a|_p$, $q_1 \equiv \gamma(1)$, then $q_1 = \exp_p(v_1^a)$. Suppose q is an arbitrary point on $\gamma(t)$, $q \equiv \gamma(t_q)$. Performing a reparametrization to $\gamma(t)$ by choosing a new parameter $t' = \alpha^{-1}t$ ($\alpha = \text{constant}$) yields a geodesic $\gamma'(t') = \gamma(t)$. Choosing an appropriate constant α we can make $\gamma'(1) = q$, and hence $q = \exp_p(v^a)$, where

$$v^a \equiv (\partial/\partial t')^a|_p = [(\partial/\partial t)^a dt/dt']_p = \alpha v_1^a.$$

Therefore, the Riemannian normal coordinate values of q are

$$x^\mu(q) = v^\mu = \alpha v_1^\mu, \quad (3.3.4)$$

where v^μ and v_1^μ are the components of v^a and v_1^a in a selected basis. $\gamma'(1) = q$ indicates that the new parameter at q has the value $t'_q = 1$, and together with the fact that $t'_q = \alpha^{-1}t_q$ due to $t' = \alpha^{-1}t$, we get $\alpha = t_q$. Hence, (3.3.4) becomes $x^\mu(q) = t_q v_1^\mu$, which can also be expressed as $x^\mu(t_q) = v_1^\mu t_q$. Since q is an arbitrary point of $\gamma(t)$, dropping the lower index q we obtain $x^\mu(t) = v_1^\mu t$; noticing that $v_1^\mu = \text{constant}$, we can see that the curve $\psi(\gamma(t))$ with parametric equations $x^\mu(t) = v_1^\mu t$ is a straight line in \mathbb{R}^n that passes through the origin. \square

Theorem 3.3.9 The Christoffel symbol of the connection ∇_a of (M, g_{ab}) in the Riemannian normal coordinate system at p satisfies $\Gamma^c{}_{ab}|_p = 0$.

Proof Any geodesic $\gamma(t)$ that passes through p can be expressed using the Riemannian normal coordinate system at p as

$$\frac{d^2 x^\mu}{dt^2} + \Gamma^\mu{}_{\nu\sigma} \frac{dx^\nu}{dt} \frac{dx^\sigma}{dt} = 0, \quad \mu = 1, \dots, n.$$

Since a Riemannian normal coordinate system (N, ψ) maps $\gamma(t)$ into a straight line in \mathbb{R}^n , we have $d^2 x^\mu / dt^2 = 0$. Thus,

$$\Gamma^\mu{}_{\nu\sigma} \frac{dx^\nu}{dt} \frac{dx^\sigma}{dt} = 0, \quad \mu = 1, \dots, n.$$

Any geodesic $\gamma(t)$ that passes through p can be expressed by the above equation. Using T^v to represent the tangent vector of the geodesic at p , then the above equation gives

$$\Gamma^\mu_{\nu\sigma}|_p T^\nu T^\sigma = 0, \quad \mu = 1, \dots, n.$$

For each μ , the left-hand side of the above equation is a quadratic polynomial with respect to n variables T^ν , and the fact that it vanishes for any T^ν renders all the coefficients being zero, i.e., $\Gamma^\mu_{\nu\sigma}|_p = 0$, $\nu, \sigma = 1, \dots, n$, and therefore $\Gamma^a_{bc}|_p = 0$. \square

[The End of Optional Reading 3.3.1]

3.4 The Riemann Curvature Tensor

3.4.1 Definition and Properties of the Riemann Curvature

A derivative operator being torsion free assures that $(\nabla_a \nabla_b - \nabla_b \nabla_a)f = 0$, i.e., $\nabla_a \nabla_b f$ is a symmetric tensor of type $(0, 2)$. Refer to the operator $\nabla_a \nabla_b - \nabla_b \nabla_a$ as the **commutator** of the derivative operator ∇_a , then ∇_a being torsion free is manifested by the fact that the action of its commutator on a function yields zero. However, the commutator of a torsion-free derivative operator acting on a tensor field of another type does not necessarily yield zero, and the Riemann curvature tensor is exactly the manifestation of this non-commutativity.

Theorem 3.4.1 Suppose $f \in \mathcal{F}$, $\omega_a \in \mathcal{F}(0, 1)$, then

$$(\nabla_a \nabla_b - \nabla_b \nabla_a)(f \omega_c) = f(\nabla_a \nabla_b - \nabla_b \nabla_a)\omega_c. \quad (3.4.1)$$

Proof Expand $\nabla_a \nabla_b(f \omega_c)$ and $\nabla_b \nabla_a(f \omega_c)$, respectively, into 4 terms and subtract them. Noticing the torsion-free condition, we get (3.4.1). \square

Theorem 3.4.2 Suppose $\omega_c, \omega'_c \in \mathcal{F}(0, 1)$ and $\omega'_c|_p = \omega_c|_p$, then

$$[(\nabla_a \nabla_b - \nabla_b \nabla_a)\omega'_c]|_p = [(\nabla_a \nabla_b - \nabla_b \nabla_a)\omega_c]|_p. \quad (3.4.2)$$

Proof Exercise 3.11. Hint: use Theorem 3.4.1. \square

Theorem 3.4.2 indicates that $(\nabla_a \nabla_b - \nabla_b \nabla_a)$ is a linear map that turns a dual vector $\omega_c|_p$ at p into a tensor $[(\nabla_a \nabla_b - \nabla_b \nabla_a)\omega_c]|_p$ of type $(0, 3)$. The way of doing this is: extend $\omega_c|_p$ arbitrarily into a dual vector field ω_c defined on a neighborhood of p , evaluate $(\nabla_a \nabla_b - \nabla_b \nabla_a)\omega_c$, and then taking the value of it at p we obtain the image of the map. Theorem 3.4.2 assures that this image does not depend on the choice of extension. Therefore, $(\nabla_a \nabla_b - \nabla_b \nabla_a)$ corresponds to a tensor of type $(1, 3)$ at p , called the **Riemann curvature tensor**, denoted by R_{abc}^d . Since p is arbitrary, R_{abc}^d is also a tensor field. Hence, we have:

Definition 1 The **Riemann curvature tensor field** R_{abc}^d of a derivative operator ∇_a is defined by the following equation

$$(\nabla_a \nabla_b - \nabla_b \nabla_a)\omega_c = R_{abc}^d \omega_d, \quad \forall \omega_c \in \mathcal{F}(0, 1). \quad (3.4.3)$$

The Riemann tensor field reflects the non-commutativity of a derivative operator, and it is a tensor field that describes the intrinsic properties of (M, ∇_a) . As long as we choose a derivative operator we can talk about its Riemann tensor. Of course, we can also talk about the Riemann tensor of a generalized Riemannian space (M, g_{ab}) , also called the Riemann tensor of g_{ab} , which is referred to as the Riemann tensor field of the derivative operator ∇_a associated with g_{ab} . A metric whose Riemann tensor field vanishes is called a **flat metric**. Now we will show that Euclidean and Minkowski metrics are both flat metrics.

Theorem 3.4.3 *The Riemann curvature tensor field of the Euclidean space $(\mathbb{R}^n, \delta_{ab})$ and Minkowski space $(\mathbb{R}^n, \eta_{ab})$ are both zero.*

Proof In Euclidean (Minkowski) space, the ordinary derivative operator ∂_a of any Cartesian (Lorentzian) system is the specific derivative operator associated with δ_{ab} . Since

$$(\partial_a \partial_b - \partial_b \partial_a)\omega_c = (dx^\mu)_a (dx^\nu)_b (dx^\sigma)_c (\partial_\mu \partial_\nu \omega_\sigma - \partial_\nu \partial_\mu \omega_\sigma) = 0, \quad \forall \omega_c,$$

the R_{abc}^d of ∂_a vanishes. \square

Therefore, Euclidean space and Minkowski space are called **flat spaces**. In fact, Minkowski space is similar to Euclidean space in many ways, and thus is also called a **pseudo-Euclidean space**.

Equation (3.4.3) reflects the non-commutativity of a derivative operator acting on a dual vector field. From this we can deduce the non-commutativity of a derivative operator acting on a tensor field of an arbitrary type $T^{c_1 \dots c_k}_{ d_1 \dots d_l}$, i.e., express $(\nabla_a \nabla_b - \nabla_b \nabla_a)T^{c_1 \dots c_k}_{ d_1 \dots d_l}$ in terms of R_{abc}^d . We have the following theorems:

Theorem 3.4.4

$$(\nabla_a \nabla_b - \nabla_b \nabla_a)v^c = -R_{abd}^c v^d \quad \forall v^c \in \mathcal{F}(1, 0). \quad (3.4.4)$$

Proof $\forall \omega_c \in \mathcal{F}(0, 1)$, we have $v^c \omega_c \in \mathcal{F}$; hence, it follows from the torsion-free condition that

$$\begin{aligned} 0 &= (\nabla_a \nabla_b - \nabla_b \nabla_a)(v^c \omega_c) = \nabla_a(v^c \nabla_b \omega_c + \omega_c \nabla_b v^c) - \nabla_b(v^c \nabla_a \omega_c + \omega_c \nabla_a v^c) \\ &= v^c \nabla_a \nabla_b \omega_c + \omega_c \nabla_a \nabla_b v^c - v^c \nabla_b \nabla_a \omega_c - \omega_c \nabla_b \nabla_a v^c. \end{aligned}$$

Thus, $\omega_c(\nabla_a \nabla_b - \nabla_b \nabla_a)v^c = -v^c(\nabla_a \nabla_b - \nabla_b \nabla_a)\omega_c = -v^c R_{abd}^c \omega_d = -\omega_c R_{abd}^c v^d$, and therefore we get (3.4.4). \square

Theorem 3.4.5 $\forall T^{c_1 \dots c_k}{}_{d_1 \dots d_l} \in \mathcal{F}(k, l)$ we have

$$(\nabla_a \nabla_b - \nabla_b \nabla_a) T^{c_1 \dots c_k}{}_{d_1 \dots d_l} = - \sum_{i=1}^k R_{abe}{}^{c_i} T^{c_1 \dots e \dots c_k}{}_{d_1 \dots d_l} + \sum_{j=1}^l R_{abd_j}{}^e T^{c_1 \dots c_k}{}_{d_1 \dots e \dots d_l}. \quad (3.4.5)$$

Proof Omitted. \square

Theorem 3.4.6 A Riemann curvature tensor has the following properties [NB: (1) and (4) are general, (2), (3) and (5) require the torsion-free condition] :

$$(1) R_{abc}{}^d = -R_{bac}{}^d; \quad (3.4.6)$$

$$(2) R_{[abc]}{}^d = 0 \text{ [cyclic identity];} \quad (3.4.7)$$

$$(3) \nabla_{[a} R_{bc]d}{}^e = 0 \text{ [Bianchi identity, published by L. Bianchi in 1902];} \quad (3.4.8)$$

if there is a metric field g_{ab} on M and $\nabla_a g_{bc} = 0$, then we can define $R_{abcd} \equiv g_{de} R_{abc}{}^e$, which also satisfies

$$(4) R_{abcd} = -R_{abdc}; \quad (3.4.9)$$

$$(5) R_{abcd} = R_{cdab}. \quad (3.4.10)$$

Proof (1) It is obvious by definition.

(2) Since $R_{[abc]}{}^d \omega_d = \nabla_{[a} \nabla_b \omega_c] - \nabla_{[b} \nabla_a \omega_c] = 2\nabla_{[a} \nabla_b \omega_c]$, to prove (3.4.7) all we have to show is that

$$\nabla_{[a} \nabla_b \omega_c] = 0, \quad \forall \omega_c \in \mathcal{F}(0, 1). \quad (3.4.11)$$

It follows from (3.1.8) that

$$\begin{aligned} \nabla_a (\nabla_b \omega_c) &= \partial_a (\nabla_b \omega_c) - \Gamma^d{}_{ab} \nabla_d \omega_c - \Gamma^d{}_{ac} \nabla_b \omega_d \\ &= \partial_a (\partial_b \omega_c - \Gamma^e{}_{bc} \omega_e) - \Gamma^d{}_{ab} \nabla_d \omega_c - \Gamma^d{}_{ac} \nabla_b \omega_d \\ &= (\partial_a \partial_b \omega_c - \Gamma^e{}_{bc} \partial_a \omega_e - \omega_e \partial_a \Gamma^e{}_{bc}) - \Gamma^d{}_{ab} \nabla_d \omega_c - \Gamma^d{}_{ac} \nabla_b \omega_d, \end{aligned} \quad (3.4.12)$$

and hence

$$\nabla_{[a} \nabla_b \omega_c] = \partial_{[a} \partial_b \omega_{c]} - \Gamma^e{}_{[bc} \partial_{a]} \omega_e - \omega_e \partial_{[a} \Gamma^e{}_{bc]} - \Gamma^d{}_{[ab} \nabla_{|d|} \omega_{c]} - \Gamma^d{}_{[ac} \nabla_b] \omega_d,$$

where $|d|$ in the lower indices $[ab|d|c]$ indicates that d does not participate in the antisymmetrization. Noticing that $\partial_a \partial_b \omega_c = \partial_b \partial_a \omega_c$ and $\Gamma^e{}_{bc} = \Gamma^e{}_{cb}$, we see from Theorem 2.6.2 (c) that each term on the right-hand side of the above equation vanishes.

(3) To prove (3.4.8), we only have to show that $\omega_e \nabla_{[a} R_{bc]d}^e = 0 \forall \omega_e \in \mathcal{F}(0, 1)$. Since

$$\begin{aligned}\omega_e \nabla_a R_{bcd}^e &= \nabla_a (R_{bcd}^e \omega_e) - R_{bcd}^e \nabla_a \omega_e \\ &= \nabla_a (\nabla_b \nabla_c \omega_d - \nabla_c \nabla_b \omega_d) - R_{bcd}^e \nabla_a \omega_e,\end{aligned}$$

one has

$$\begin{aligned}\omega_e \nabla_{[a} R_{bc]d}^e &= \nabla_{[a} \nabla_b \nabla_{c]} \omega_d - \nabla_{[a} \nabla_c \nabla_{b]} \omega_d - R_{[bc|d]}^e \nabla_{a]} \omega_e \\ &= \nabla_{[a} \nabla_b \nabla_{c]} \omega_d - \nabla_{[b} \nabla_a \nabla_{c]} \omega_d - R_{[bc|d]}^e \nabla_{a]} \omega_e.\end{aligned}\quad (3.4.13)$$

To derive the sum of the first two terms on the right, first we write out the expression without the square bracket

$$\nabla_a \nabla_b \nabla_c \omega_d - \nabla_b \nabla_a \nabla_c \omega_d = (\nabla_a \nabla_b - \nabla_b \nabla_a) \nabla_c \omega_d = R_{abc}^e \nabla_e \omega_d + R_{abd}^e \nabla_c \omega_e,$$

where in the second equality we used (3.4.5). Antisymmetrizing the lower indices a, b, c , and noticing (3.4.7), then we have

$$\nabla_{[a} \nabla_b \nabla_{c]} \omega_d - \nabla_{[b} \nabla_a \nabla_{c]} \omega_d = R_{[ab|d]}^e \nabla_{c]} \omega_e = R_{[bc|d]}^e \nabla_{a]} \omega_e,$$

which indicates that the right-hand side of (3.4.13) vanishes. Therefore, $\omega_e \nabla_{[a} R_{bc]d}^e = 0$.

(4) Applying (3.4.5) to g_{cd} , it follows from $\nabla_a g_{cd} = 0$ that

$$0 = (\nabla_a \nabla_b - \nabla_b \nabla_a) g_{cd} = R_{abc}^e g_{ed} + R_{abd}^e g_{ce} = R_{abcd} + R_{abdc},$$

and hence (3.4.9) holds.

(5) Left as an exercise (Exercise 3.12). □

Remark 1 Suppose $\dim M = n$, then R_{abcd} has in total n^4 components $R_{\mu\nu\sigma\rho}$. However, since the algebraic equations (3.4.6), (3.4.7), (3.4.9) and (3.4.10) are satisfied, the number of independent components is only [for a proof, see Bergmann (1976) pp. 172–174]

$$N = \frac{n^2(n^2 - 1)}{12}.$$

After a metric is chosen, each tensor T_{ab} of type $(0, 2)$ corresponds to a tensor $T^a{}_b \equiv g^{ac} T_{cb}$ of type $(1, 1)$, which is nothing but a linear transformation on a vector space. The components of this linear transformation in an arbitrary basis form a matrix, and the matrices in different bases are similar to each other; hence, they have the same trace, whose value is $T^a{}_a = g^{ac} T_{ac}$, called the trace of the tensor $T^a{}_b$, also called the trace of T_{ab} . Similarly, for a given tensor R_{abcd} of type $(0, 4)$, we can in principle obtain the following six “traces” through contraction [each “trace” is a tensor of type $(0, 2)$]: $g^{ab} R_{abcd}$, $g^{ac} R_{abcd}$, $g^{ad} R_{abcd}$, $g^{bc} R_{abcd}$, $g^{bd} R_{abcd}$, $g^{cd} R_{abcd}$.

However, due to the properties of R_{abcd} which comes from lowering the upper index of the Riemann tensor R_{abc}^d [(1), (4), (5) of Theorem 3.4.6] and the symmetry of g^{ac} , it is easy to see from (d) of Theorem 2.6.2 that the first and the sixth contractions above vanishes; the second and the fifth are equal (reason: $g^{ac}R_{abcd}=g^{ac}R_{badc}$, which is essentially the same as $g^{bd}R_{abcd}$, we do not write $g^{ac}R_{abcd}=g^{bd}R_{abcd}$ only because we need to take care of the balance of indices); the third and the forth are equal and they are the negative of the second and the fifth ones. Hence, among these six contractions there is only a single independent one, we can take, for example, $g^{bd}R_{abcd}$, denoted by R_{ac} , called the **Ricci tensor**. What should be emphasized is that we do not need a metric to define the Ricci tensor since $R_{ac}\equiv R_{abc}^b$ is endowed with a clear meaning. We can also take the trace of R_{ac} using the metric, i.e., $g^{ac}R_{ac}$, denoted by R , called the **scalar curvature**. From (3.4.10), it is easy to show that $R_{ac}=R_{ca}$. Besides, one should also be acquainted with the traceless part of R_{abc}^d , which is called the **Weyl tensor**, defined as follows:

Definition 2 For a generalized Riemannian space of dimension $n \geq 3$, the **Weyl tensor** C_{abcd} is defined by the following expression:

$$C_{abcd} := R_{abcd} - \frac{2}{n-2}(g_{a[c}R_{d]b} - g_{b[c}R_{d]a}) + \frac{2}{(n-1)(n-2)}Rg_{a[c}g_{d]b}. \quad (3.4.14)$$

Theorem 3.4.7 Weyl tensors have the following properties:

- (1) $C_{abcd} = -C_{bacd} = -C_{abdc} = C_{cdab}, \quad C_{[abc]d} = 0.$ (3.4.15)
- (2) The trace of C_{abcd} over any pair of indices vanishes, e.g., $g^{ac}C_{abcd} = 0$.

Proof Exercise. \square

Remark 2 Equation (3.4.14) indicates that R_{abcd} is the summation of its traceless part C_{abcd} and its trace part

$$\frac{2}{n-2}(g_{a[c}R_{d]b} - g_{b[c}R_{d]a}) - \frac{2}{(n-1)(n-2)}Rg_{a[c}g_{d]b}.$$

Definition 3 The **Einstein tensor** of a generalized Riemannian space is defined by

$$G_{ab} := R_{ab} - \frac{1}{2}Rg_{ab}. \quad (3.4.16)$$

Theorem 3.4.8

$$\nabla^a G_{ab} = 0 \text{ (where } \nabla^a G_{ab} \equiv g^{ac}\nabla_c G_{ab}). \quad (3.4.17)$$

Proof From the Bianchi identity (3.4.8) and (3.4.6) we have $0 = \nabla_a R_{bcd}^e + \nabla_c R_{abd}^e + \nabla_b R_{cad}^e$. Contracting indices a and e yields $0 = \nabla_a R_{bcd}^a + \nabla_c R_{abd}^a + \nabla_b R_{cad}^a = \nabla_a R_{bcd}^a - \nabla_c R_{bd}^a + \nabla_b R_{cd}^a$. Acting g^{bd} on it we get

$$\begin{aligned} 0 &= g^{bd} \nabla_a R_{bcd}^a - g^{bd} \nabla_c R_{bd} + g^{bd} \nabla_b R_{cd} \\ &= \nabla_a R_c^a - \nabla_c R + \nabla_b R_c^b = 2\nabla_a R_c^a - \nabla_c R. \end{aligned} \quad (3.4.18)$$

Hence, $\nabla^a G_{ab} = \nabla^a R_{ab} - \frac{1}{2}g^{ab}\nabla^a R = \nabla_a R_b^a - \frac{1}{2}\nabla_b R = 0$, where we used $R_{ab} = R_{ba}$ in the second equality and (3.4.18) in the third equality. \square

Equation (3.4.17) that the Einstein tensor satisfies is significant for establishing Einstein's equation of general relativity, for details, see Sect. 7.7.

3.4.2 Computing Riemann Curvature from a Metric

Suppose M has a given metric g_{ab} , from $\nabla_a g_{bc} = 0$ a unique connection ∇_a is determined, and thus we have a Riemann tensor R_{abc}^d . A common problem is to compute R_{abc}^d from the given g_{ab} . Computing a tensor means deriving its components in a certain basis. There are two types of basis: coordinate basis and non-coordinate basis. In this section, we only talk about the method of computing curvature using a coordinate basis; the methods using non-coordinate bases are introduced in Sects. 5.7 and 8.7.

After we choose an arbitrary coordinate system, the components $g_{\mu\nu}$ of the metric are then known, and in this coordinate system the connection ∇_a satisfying $\nabla_a g_{bc} = 0$ can be characterized by its Christoffel symbol in this system:

$$\Gamma^\sigma_{\mu\nu} = \frac{1}{2}g^{\sigma\rho}(g_{\rho\mu,\nu} + g_{\nu\rho,\mu} - g_{\mu\nu,\rho}) \quad [\text{i.e., (3.2.10')}] . \quad (3.4.19)$$

$\Gamma^\sigma_{\mu\nu}$ has three component indices, and thus $\{\Gamma^\sigma_{\mu\nu}\}$ contains n^3 numbers. The symmetry $\Gamma^\sigma_{\mu\nu} = \Gamma^\sigma_{\nu\mu}$ makes it so that only $n^2(n+1)/2$ among the n^3 numbers are independent (when $n = 4$ there are 40 independent numbers). The first step for the calculation is to derive all the nonvanishing $\Gamma^\sigma_{\mu\nu}$ from the given $g_{\mu\nu}$.

From the definition of the Riemann tensor we have $R_{abc}^d \omega_d = 2\nabla_{[a} \nabla_{b]} \omega_c$, where $\nabla_a \nabla_b \omega_c$ can be expressed in six terms using (3.4.12) (there are five terms in this equation, and the fifth term can be expanded into two terms, i.e., $\partial_b \omega_d - \Gamma^e{}_{bd} \omega_e$). Antisymmetrizing the indices a, b in each term, and noting that $\partial_{[a} \partial_{b]} \omega_c = 0$, $\Gamma^d{}_{[ab]} = \Gamma^d{}_{(ab)} = 0$, we obtain

$$\begin{aligned} R_{abc}^d \omega_d &= 2(-\Gamma^e{}_{c[b} \partial_{a]} \omega_e - \omega_e \partial_{[a} \Gamma^e{}_{b]c} - \Gamma^d{}_{c[a} \partial_{b]} \omega_d + \Gamma^d{}_{c[a} \Gamma^e{}_{b]d} \omega_e) \\ &= -2\omega_d \partial_{[a} \Gamma^d{}_{b]c} + 2\Gamma^e{}_{c[a} \Gamma^d{}_{b]e} \omega_d, \quad \forall \omega_d \in \mathcal{F}(0, 1). \end{aligned}$$

Hence,

$$R_{abc}^d = -2\partial_{[a} \Gamma^d{}_{b]c} + 2\Gamma^e{}_{c[a} \Gamma^d{}_{b]e}, \quad (3.4.20)$$

whose coordinate components are

$$R_{\mu\nu\sigma}{}^\rho = \Gamma^\rho{}_{\mu\sigma,\nu} - \Gamma^\rho{}_{\nu\sigma,\mu} + \Gamma^\lambda{}_{\sigma\mu}\Gamma^\rho{}_{\nu\lambda} - \Gamma^\lambda{}_{\sigma\nu}\Gamma^\rho{}_{\mu\lambda}, \quad (3.4.20')$$

where $\Gamma^\rho{}_{\mu\sigma,\nu} \equiv \partial\Gamma^\rho{}_{\mu\sigma}/\partial x^\nu$. From the equation above we can also obtain the expression for the coordinate components of the Ricci tensor

$$R_{\mu\sigma} = R_{\mu\nu\sigma}{}^\nu = \Gamma^\nu{}_{\mu\sigma,\nu} - \Gamma^\nu{}_{\nu\sigma,\mu} + \Gamma^\lambda{}_{\mu\sigma}\Gamma^\nu{}_{\lambda\nu} - \Gamma^\lambda{}_{\nu\sigma}\Gamma^\nu{}_{\lambda\mu}. \quad (3.4.21)$$

[Optional Reading 3.4.1]

If the components $g_{\mu\nu}$ of a metric g_{ab} in a coordinate system are all constants, then all of its Christoffel symbols vanish ($\Gamma^\sigma{}_{\mu\nu} = 0$), and from (3.4.20') we know that its Riemann tensor $R_{abc}{}^d = 0$; thus, (at least in this coordinate patch) it is a flat metric. Conversely, if we know that for a g_{ab} there is $R_{abc}{}^d = 0$, does there always exist a coordinate system such that the coordinate components $g_{\mu\nu}$ of g_{ab} are all constants? The answer is affirmative. See the following theorem.

Theorem 3.4.9 *A metric field g_{ab} is (locally) flat (i.e., $R_{abc}{}^d = 0$) if and only if there exists a coordinate system such that the coordinate components of g_{ab} are all constants.*

Proof The proof of this theorem requires techniques that we have not covered yet, see Appendix J of Volume III. \square

[The End of Optional Reading 3.4.1]

[Optional Reading 3.4.2]

Equation (3.4.12) contains $\Gamma^\nu{}_{\nu\sigma}$. In fact, “contracted Christoffel symbols” like this are involved in many computations. For instance, inspired by the definition of the divergence $\vec{\nabla} \cdot \vec{v}$ in a 3-dimensional Euclidean space, we can define the divergence of a vector field v^a in (M, ∇_a) as $\nabla_a v^a$ (we may also call $\nabla_a T^{ab}$ the divergence of a tensor field T^{ab}). Since $\nabla_a v^a = \partial_a v^a + \Gamma^a{}_{ab} v^b$, we need to deal with the “contracted Christoffel symbol” $\Gamma^a{}_{ab}$ when calculating the divergence. Now we shall derive the expression for $\Gamma^\nu{}_{\nu\sigma}$. It follows from (3.4.19) that

$$\Gamma^\mu{}_{\mu\sigma} = \frac{1}{2} g^{\mu\lambda} (g_{\sigma\lambda,\mu} + g_{\mu\lambda,\sigma} - g_{\mu\sigma,\lambda}) = \left(\frac{1}{2} g^{\mu\lambda} g_{\mu\lambda,\sigma} + g^{\mu\lambda} g_{\sigma[\lambda,\mu]} \right) = \frac{1}{2} g^{\mu\lambda} g_{\mu\lambda,\sigma},$$

where we used the fact that $g^{[\mu\lambda]} = 0$ in the last step. This equation can be rewritten as

$$\Gamma^\mu{}_{\mu\sigma} = \frac{1}{2} g^{\mu\lambda} \frac{\partial g_{\mu\lambda}}{\partial x^\sigma}. \quad (3.4.22)$$

On the other hand, the determinant g of the matrix constituted by $g_{\mu\lambda}$ can be expanded with respect to the μ th row as $g = g_{\mu\lambda} A^{\mu\lambda}$ (where $A^{\mu\lambda}$ is the cofactor of $g_{\mu\lambda}$, and the sum is only taken over λ); hence, $\partial g / \partial g_{\mu\lambda} = A^{\mu\lambda}$. Thus, from the expression for the inverse matrix elements $g^{\mu\lambda} = A^{\lambda\mu} / g$ we have

$$\frac{\partial g}{\partial g_{\mu\lambda}} = g g^{\mu\lambda}. \quad (3.4.23)$$

Since $g_{\mu\lambda}$ are functions of the coordinates x^σ , g is a function of x^σ as well, and

$$\frac{\partial g}{\partial x^\sigma} = \frac{\partial g}{\partial g_{\mu\lambda}} \frac{\partial g_{\mu\lambda}}{\partial x^\sigma} = g g^{\mu\lambda} \frac{\partial g_{\mu\lambda}}{\partial x^\sigma}, \quad (3.4.24)$$

where (3.4.23) is used in the last step. Combining (3.4.22) and (3.4.24) yields

$$\Gamma^\mu_{\mu\sigma} = \frac{1}{2g} \frac{\partial g}{\partial x^\sigma} = \frac{1}{\sqrt{|g|}} \frac{\partial \sqrt{|g|}}{\partial x^\sigma}. \quad (3.4.25)$$

This is the expression for the “contracted Christoffel symbol”. The divergence $\nabla_a v^a$ (as a scalar field) can be derived by means of an arbitrary basis. Using the coordinate basis, it is easy to derive from (3.4.25) and $\nabla_a v^a = \partial_a v^a + \Gamma^a_{ab} v^b$ that

$$\nabla_a v^a = \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x^\sigma} (\sqrt{|g|} v^\sigma). \quad (3.4.26)$$

As an example of the application, now we derive the expression for the divergence $\vec{\nabla} \cdot \vec{v}$ of a vector field \vec{v} in the 3-dimensional Euclidean space in both the Cartesian and the spherical coordinate systems. First, we rewrite the above equation as

$$\vec{\nabla} \cdot \vec{v} = \nabla_a v^a = \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x^i} (\sqrt{|g|} v^i). \quad (3.4.27)$$

(1) For a Cartesian coordinate system, $g = 1$, $\vec{\nabla} \cdot \vec{v} = \frac{\partial v^i}{\partial x^i} = \frac{\partial v^1}{\partial x^1} + \frac{\partial v^2}{\partial x^2} + \frac{\partial v^3}{\partial x^3}$; this is the familiar formula for divergence.

(2) For a spherical coordinate system, $\sqrt{g} = r^2 \sin^2 \theta$,

$$\begin{aligned} \vec{\nabla} \cdot \vec{v} &= \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial x^i} (v^i r^2 \sin \theta) \\ &= \frac{1}{r^2 \sin \theta} \left[\frac{\partial(v^1 r^2 \sin \theta)}{\partial r} + \frac{\partial(v^2 r^2 \sin \theta)}{\partial \theta} + \frac{\partial(v^3 r^2 \sin \theta)}{\partial \varphi} \right], \end{aligned} \quad (3.4.28)$$

where v^1, v^2, v^3 are the components of v^a in the coordinate basis $\{(\partial/\partial r)^a, (\partial/\partial \theta)^a, (\partial/\partial \varphi)^a\}$. However, normally the formula in an electrodynamics textbook is written in terms of the components of v^a in an orthonormal basis $\{(e_r)^a, (e_\theta)^a, (e_\varphi)^a\}$ (denoted by v^r, v^θ, v^φ). Note that

$$(e_r)^a = (\partial/\partial r)^a, \quad (e_\theta)^a = r^{-1} (\partial/\partial \theta)^a, \quad (e_\varphi)^a = (r \sin \theta)^{-1} (\partial/\partial \varphi)^a,$$

which means $v^1 = v^r, v^2 = r^{-1} v^\theta, v^3 = (r \sin \theta)^{-1} v^\varphi$. Plugging these into (3.4.27) yields

$$\vec{\nabla} \cdot \vec{v} = \frac{1}{r^2} \frac{\partial(v^r r^2)}{\partial r} + \frac{1}{r \sin \theta} \frac{\partial(v^\theta \sin \theta)}{\partial \theta} + \frac{1}{r \sin \theta} \frac{\partial(v^\varphi)}{\partial \varphi},$$

which agrees with the formula in electrodynamics textbooks.

[The End of Optional Reading 3.4.2]

3.5 The Intrinsic Curvature and the Extrinsic Curvature

According to our intuition, a plane is flat while a curved surface is not. More precisely, these “flat” and “curved” surfaces in our mind are all 2-dimensional surfaces (such as spherical and cylindrical surfaces) embedded in the 3-dimensional Euclidean space. Now we ask: given an n -dimensional manifold, can we talk about if it is curved by following the same idea? As long as it can be embedded into an $(n+1)$ -dimensional manifold, the answer will be yes. The curvature defined by embedding a manifold in

a manifold with one extra dimension is called the “extrinsic curvature”, which has a precise definition (for details see Chap. 14). According to this definition, both of a sphere and a cylindrical surface in 3-dimensional Euclidean space have a nonzero curvature, which tallies with our intuition. However, the Riemann curvature we introduced in this chapter is the intrinsic curvature, which reflects the “intrinsic warping” of a manifold M after a connection ∇_a is assigned. Unlike the extrinsic curvature, there is no need to embed M in a one-higher dimensional manifold to tell the intrinsic curvature. [Generally speaking, any property of (M, g_{ab}) that can be determined by just g_{ab} (without having to embed the manifold in a higher dimensional manifold) is called an **intrinsic property** of (M, g_{ab}) .] The term “intrinsic curvature” actually just reflects the following three equivalent properties; a generalized Riemannian space with these properties is called a curved space.

(1) The non-commutativity of the derivative operator, i.e., $(\nabla_a \nabla_b - \nabla_b \nabla_a)\omega_c = R_{abc}{}^d \omega_d$, $\forall \omega_c \in \mathcal{F}(0, 1)$, where the nonvanishing tensor field $R_{abc}{}^d$ is used as the definition of the intrinsic (Riemann) curvature, see Sect. 3.4.

(2) The curve-dependence of the parallel transport of a vector.

As we have discussed in Sect. 3.2, for two points p and q in (M, ∇_a) , there exists a curve-dependent translation map between their tangent spaces V_p and V_q ; that is, for a curve between p and q , any vector v^a at p determines a vector field \tilde{v}^a (satisfies $\tilde{v}^a|_p = v^a$) parallelly transported along the curve whose value at q can be defined as the image of v^a . In other words, $\tilde{v}^a|_q$ is the result of v^a parallelly transported to q . For Euclidean, Minkowski and any other flat space, this parallel transport is curve-independent; thus, there is no need to specify a curve when we talk about “parallelly transporting a vector at p to q ”. This simplicity is called the absoluteness of the parallel transport, which is pretty familiar to us. (Do you specify a curve when you parallel transport a vector from a point to another point in Euclidean space?) However, it is not as simple for a curved space. It can be proved that [see Wald (1984) pp. 37–38; Straumann (1984) Theorem 5.7] a necessary and sufficient condition for the intrinsic curvature $R_{abc}{}^d$ to be nonvanishing is that there exists a closed curve such that a vector at a point on the curve will not return to itself when parallelly transported along the curve; therefore, the parallel transport depends on a curve (there is only a curve-dependent concept of parallel transport). Spherical geometry provides a simple but intuitive example of this phenomenon:

Example 1 It can be computed that the $R_{abc}{}^d$ of a 2-dimensional sphere (together with the induced metric) in a 3-dimensional Euclidean space is nonvanishing (see Exercise 3.13). Figure 3.10 indicates that there exists a closed curve $abca$ (each segment is an arc of a great circle) such that a vector fails to return to itself when parallelly transported along the curve. Take the vector v^a at a in the figure for example; it is a tangent vector of a geodesic ab . Since the tangent of a geodesic is parallelly transported along the geodesic, the result of the parallel transport of v^a to b is u_a (see Fig. 3.10), which is orthogonal to the tangent vector T^a of bc . Since the parallel transport preserves the orthogonality, as shown in the figure, the result of u^a parallelly transported to c is w^a . w^a is tangent to a geodesic ac , and hence v^a coming from w^a parallelly transported to c should also be tangent to ac ; therefore, $v^a \neq w^a$.

Fig. 3.10 A vector v^a at a becomes $v'^a \neq v^a$ after being parallelly transported along a closed curve $abca$ on the sphere

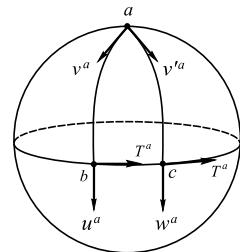
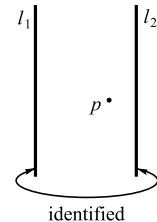


Fig. 3.11 Identifying l_1 and l_2 yields a cylindrical surface



(3) There exist geodesics that are parallel at first which become not parallel.

The two meridians in Fig. 3.10 give an intuitive example. For the precise meaning see Sect. 7.6.

The curvature tensor field $R_{abc}{}^d$ of a flat space vanishes, and thus it does not have any of the three properties above. Specifically, ① the derivative operator ∂_a associated with the flat metric (i.e., the ordinary derivative operator of a Cartesian or Lorentzian system) does not have the non-commutativity; ② the parallel transport of a vector does not depend on the curve, and thus one can talk about the “absolute parallel transport” of a vector; ③ parallel lines will never intersect.

The intrinsic curvature and the extrinsic curvature are two different concepts. For instance, a 2-dimensional cylindrical surface in 3-dimensional Euclidean space has a nonzero extrinsic curvature but a zero intrinsic curvature. A cylindrical surface can be viewed as the part between two parallel lines l_1 and l_2 on a plane after identifying (gluing together) these two lines (see Fig. 3.11). Since the computation of $R_{abc}{}^d$ at p only involves a neighborhood of p , it would not become nonzero due to the identification of l_1 and l_2 .

Exercises

~3.1. Now we give up the torsion-free condition,

(1) Show that there exists a tensor $T^c{}_{ab}$ (called the **torsion tensor**) such that

$$\nabla_a \nabla_b f - \nabla_b \nabla_a f = -T^c{}_{ab} \nabla_c f, \quad \forall f \in \mathcal{F}.$$

Hint: set $\tilde{\nabla}_a$ as a torsion-free operator, imitate the derivation in Theorem 3.1.4.

- (2) Show that $T^c{}_{ab}u^av^b = u^a\nabla_a v^c - v^a\nabla_a u^c - [u, v]^c \quad \forall u^a, v^a \in \mathcal{F}(1, 0)$.
- 3.2. Suppose v^a is a vector field, v^ν and v'^ν are the components of v^a in coordinate systems $\{x^\nu\}$ and $\{x'^\nu\}$, respectively; $A^\nu{}_\mu \equiv \partial v^\nu / \partial x^\mu$, $A'^\nu{}_\mu \equiv \partial v'^\nu / \partial x'^\mu$. Show that the relation between $A^\nu{}_\mu$ and $A'^\nu{}_\mu$ generally does not satisfy the tensor components transformation law. Hint: use the transformation law between v^ν and v'^ν .
- ~3.3. Prove Theorem 3.1.7.
- 3.4. Using the following definition of $\Gamma^\sigma{}_{\mu\nu} : \left(\frac{\partial}{\partial x^\nu}\right)^b \nabla_b \left(\frac{\partial}{\partial x^\mu}\right)^a = \Gamma^\sigma{}_{\mu\nu} \left(\frac{\partial}{\partial x^\sigma}\right)^a$, show that
- $\Gamma^\sigma{}_{\mu\nu} = \Gamma^\sigma{}_{\nu\mu}$; (Hint: use the fact that ∇_a is torsion free and that coordinate basis vectors commute.)
 - $v^\nu{}_{;\mu} = v^\nu{}_{,\mu} + \Gamma^\nu{}_{\mu\beta}v^\beta$. (NB: This is actually an equivalent definition of Christoffel symbols.)
- ~3.5. Determine whether each of the equations are true or false:
- $\nabla_a(dx^\mu)_b = 0$;
 - $v^\nu{}_{;\mu} = (\nabla_a v^b)(\partial/\partial x^\mu)^a (dx^\nu)_b$;
 - $v^\nu{}_{,\mu} = (\partial_a v^b)(\partial/\partial x^\mu)^a (dx^\nu)_b$;
 - $v^\nu{}_{;\mu} = (\partial/\partial x^\mu)^a \nabla_a v^\nu$;
 - $v^\nu{}_{,\mu} = (\partial/\partial x^\mu)^a \nabla_a v^\nu$.
- ~3.6. Suppose $C(t)$ is a curve in the coordinate patch of $\{x^\mu\}$, $x^\mu(t)$ is the parametric representation of $C(t)$ in this coordinate system, and v^a is a vector field on $C(t)$. Let $Dv^\mu/dt \equiv (dx^\mu)_a(\partial/\partial t)^b \nabla_b v^a$. Show that
- $$\frac{Dv^\mu}{dt} = \frac{dv^\mu}{dt} + \Gamma^\mu{}_{\nu\sigma} v^\sigma \frac{dx^\nu(t)}{dt}.$$
- ~3.7. Find all of the nonvanishing $\Gamma^\sigma{}_{\mu\nu}$ of a spherical coordinate system in the 3-dimensional Euclidean space.
- 3.8 Suppose I is an interval of \mathbb{R} , and $C : I \rightarrow M$ is a curve in (M, ∇_a) . Show that $\forall s, t \in I$, the translation map $\psi : V_{C(s)} \rightarrow V_{C(t)}$ (see Fig. 3.2) is an isomorphism.
- ~3.9. Prove Theorems 3.3.2, 3.3.3 and 3.3.5.
- ~3.10. (a) Write down the geodesic equation of the spherical metric $ds^2 = R^2(d\theta^2 + \sin^2\theta d\varphi^2)$ (where R is a constant); (b) verify that any arc of a great circle satisfies the geodesic equation. Hint: choose a spherical coordinate system $\{\theta, \varphi\}$ such that the given great circle arc is part of the equator, and use φ as the affine parameter.
- ~3.11. Prove Theorem 3.4.2.
- *3.12. Prove (3.4.10).

- ~3.13. Derive all of the components of the Riemann tensor of the spherical metric (see Exercise 3.10) in the $\{\theta, \varphi\}$ coordinate system.
- 3.14. Derive all of the components of the Riemann tensor of the metric $ds^2 = \Omega^2(t, x)(-dt^2 + dx^2)$ in the $\{t, x\}$ coordinate system (use $\dot{\Omega}$ and Ω' to represent the partial derivatives of the function Ω with respect to t and x , respectively).
- 3.15. Derive all of the components of the Riemann tensor of the metric $ds^2 = z^{-1/2}(-dt^2 + dz^2) + z(dx^2 + dy^2)$ in the $\{t, x, y, z\}$ coordinate system.
- 3.16. Suppose $\alpha(z)$, $\beta(z)$, $\gamma(z)$ are three arbitrary functions, $h = t + \alpha(z)x + \beta(z)y + \gamma(z)$. Derive all of the components of the Riemann tensor of the metric

$$ds^2 = -dt^2 + dx^2 + dy^2 + h^2 dz^2$$

in the $\{t, x, y, z\}$ coordinate system.

- 3.17. Show that the Einstein tensor of a 2-dimensional generalized Riemannian space vanishes. Hint: the Riemann tensor of a 2-dimensional generalized Riemannian space has only one independent component.

References

- Bergmann, P. G. (1976), *Introduction to the Theory of Relativity*, Dover Publications INC, New York.
- Chern, S. S., Chen, W. & Lam, K. S. (1999), *Lectures on Differential Geometry*, World Scientific Publishing Company, Singapore.
- Hawking, S. W. & Ellis, G. F. R. (1973), *The Large Scale Structure of Space-Time*, Cambridge University Press, Cambridge.
- Straumann, N. (1984), *General Relativity and Relativistic Astrophysics*, Springer-Verlag, Berlin.
- Wald, R. M. (1984), *General Relativity*, The University of Chicago Press, Chicago.

Chapter 4

Lie Derivatives, Killing Fields and Hypersurfaces



4.1 Maps of Manifolds

Suppose M and N are manifolds (whose dimensions can be different) and $\phi : M \rightarrow N$ is a smooth map. Let $\mathcal{F}_M(k, l)$ and $\mathcal{F}_N(k, l)$ represent the collection of all smooth tensor fields of type (k, l) on M and N , respectively. ϕ naturally induces a series of maps as follows.

Definition 1 The **pullback map** $\phi^* : \mathcal{F}_N \rightarrow \mathcal{F}_M$ is defined as

$$(\phi^* f)|_p := f|_{\phi(p)}, \quad \forall f \in \mathcal{F}_N, p \in M,$$

i.e., $\phi^* f = f \circ \phi$, see Fig. 4.1.

From Definition 1 it is not difficult to prove that

- (1) $\phi^* : \mathcal{F}_N \rightarrow \mathcal{F}_M$ is a linear map, i.e.,
$$\phi^*(\alpha f + \beta g) = \alpha \phi^*(f) + \beta \phi^*(g) \quad \forall f, g \in \mathcal{F}_N, \quad \alpha, \beta \in \mathbb{R}.$$
- (2) $\phi^*(fg) = \phi^*(f)\phi^*(g), \quad \forall f, g \in \mathcal{F}_N.$ (4.1.1)

Definition 2 For any point in M one can define the **pushforward map** $\phi_* : V_p \rightarrow V_{\phi(p)}$ as follows: $\forall v^a \in V_p$, define its image $\phi_* v^a \in V_{\phi(p)}$ as

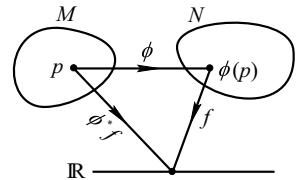
$$(\phi_* v)(f) := v(\phi^* f), \quad \forall f \in \mathcal{F}_N. \quad (4.1.2)$$

It should also be verified (Exercise 4.1) that the $\phi_* v^a$ defined in this manner satisfies the two conditions for a vector in Definition 2 of Sect. 2.2 and is thus indeed a vector at $\phi(p)$. Many works refer to ϕ_* as the **tangent map** of ϕ .

Theorem 4.1.1 $\phi_* : V_p \rightarrow V_{\phi(p)}$ is a linear map, i.e.,

$$\phi_*(\alpha u^a + \beta v^a) = \alpha \phi_* u^a + \beta \phi_* v^a, \quad \forall u^a, v^a \in V_p, \quad \alpha, \beta \in \mathbb{R}.$$

Fig. 4.1 The definition of $\phi^* f$



Proof Exercise 4.2. □

Theorem 4.1.2 Suppose $C(t)$ is a curve in M and T^a is the tangent vector of the curve at a point $C(t_0)$, then $\phi_* T^a \in V_{\phi(C(t_0))}$ is the tangent vector of the curve $\phi(C(t))$ at $\phi(C(t_0))$ (the image of the tangent vector of a curve is the tangent vector of the image of the curve).

Proof Exercise 4.2. Hint: use the definition of the tangent vector of a curve [see (2.2.6)]. □

Definition 3 The pullback map can be extended to $\phi^* : \mathcal{F}_N(0, l) \rightarrow \mathcal{F}_M(0, l)$ in the following way: $\forall T \in \mathcal{F}_N(0, l)$ define $\phi^* T \in \mathcal{F}_M(0, l)$ as

$$(\phi^* T)_{a_1 \dots a_l}|_p(v_1)^{a_1} \dots (v_l)^{a_l} := T_{a_1 \dots a_l}|_{\phi(p)}(\phi_* v_1)^{a_1} \dots (\phi_* v_l)^{a_l}, \\ \forall p \in M, \quad v_1, \dots, v_l \in V_p. \quad (4.1.3)$$

Definition 4 $\forall p \in M$ the pushforward map can be extended to $\phi_* : \mathcal{T}_{V_p}(k, 0) \rightarrow \mathcal{T}_{V_{\phi(p)}}(k, 0)$ in the following manner [namely, ϕ_* is a map that turns a tensor of type $(k, 0)$ at p into a tensor of the same type at $\phi(p)$]: $\forall T \in \mathcal{T}_{V_p}(k, 0)$ its image $\phi_* T \in \mathcal{T}_{V_{\phi(p)}}(k, 0)$ is defined by the following equation:

$$(\phi_* T)^{a_1 \dots a_k}(\omega^1)_{a_1} \dots (\omega^k)_{a_k} := T^{a_1 \dots a_k}(\phi^* \omega^1)_{a_1} \dots (\phi^* \omega^k)_{a_k}, \\ \forall \omega_1, \dots, \omega_k \in V_{\phi(p)}^*,$$

where $(\phi^* \omega)_a$ is defined as $(\phi^* \omega)_a v^a := \omega_a(\phi_* v)^a \quad \forall v^a \in V_p$.

Remark 1 Definition 2 is nothing but a special case of Definition 4 when $k = 1$. If we refer to a scalar field as a tensor field of type $(0, 0)$, then Definition 1 is nothing but a special case of Definition 3 when $l = 0$. Definition 3 indicates that the pullback map ϕ^* can turn a tensor field of type $(0, l)$ on N into a tensor field of the same type on M , and thus is a map that turns a *field* into a *field*; while according to Definition 4, the pushforward map ϕ_* only turns a tensor of type $(k, 0)$ at a point p in M into a tensor of the same type at the image point $\phi(p)$. Can we extend ϕ_* to a map that turns a tensor *field* of type $(k, 0)$ on M into a tensor *field* of the same type on N ? Generally speaking, we cannot. Take a vector field as an example. Given a vector field v on M , to define the image field $\phi_* v$ on N we need to define a vector for each point of N , which is always related to the inverse image $\phi^{-1}(q)$. [This is analogous

to Definition 3, according to which ϕ^* can turn a field T on N into a field ϕ^*T on M , and to define the value ϕ^*T at any point of M one will need the value of T at $\phi(p)$.] If ϕ is not an onto map, then $\phi^{-1}(q)$ may not exist, and thus we cannot use the v at $\phi^{-1}(q)$ as the v on the right-hand side of (4.1.2). If ϕ is not a one-to-one map, then there may be more than one point for the inverse image $\phi^{-1}(q)$, and thus we cannot determine the v of which inverse point should be defined as the v on the right-hand side of (4.1.2). This implies that if ϕ is just a smooth map, then ϕ_* cannot necessarily pushforward a field to a field. However, if $\phi : M \rightarrow N$ is a diffeomorphism, then the trouble we just mentioned will disappear. The pushforward map ϕ_* can be viewed as a map that turns a tensor field of type $(k, 0)$ on M into a tensor field of the same type on N , i.e., $\phi_* : \mathcal{F}_M(k, 0) \rightarrow \mathcal{F}_N(k, 0)$. Furthermore, since ϕ^{-1} exists and is smooth, its pullback map ϕ^{-1*} maps $\mathcal{F}_M(0, l)$ to $\mathcal{F}_N(0, l)$, which can be regarded as the pushforward map of ϕ . Therefore, ϕ_* can be generalized even further as $\phi_* : \mathcal{F}_M(k, l) \rightarrow \mathcal{F}_N(k, l)$. For instance, suppose $T^a{}_b \in \mathcal{F}_M(1, 1)$, then $(\phi_* T)^a{}_b \in \mathcal{F}_N(1, 1)$ is defined as

$$(\phi_* T)^a{}_b|_q \omega_a v^b := T^a{}_b|_{\phi^{-1}(q)} (\phi^* \omega)_a (\phi^* v)^b, \quad \forall q \in N, \omega_a \in V_q^*, v \in V_q,$$

where $(\phi^* v)^b$ should be understood as $(\phi_*^{-1} v)^b$. Similarly, the pullback map can also be generalized as $\phi^* : \mathcal{F}_N(k, l) \rightarrow \mathcal{F}_M(k, l)$. The generalized ϕ_* and ϕ^* are still linear maps and are the inverse of each other.

Suppose $\phi : M \rightarrow N$ is a diffeomorphism, $p \in M$, $\{x^\mu\}$ and $\{y^\mu\}$ are local coordinate systems of M and N , respectively, whose coordinate patches O_1 and O_2 satisfy $p \in O_1$ and $\phi(p) \in O_2$. Thus, $p \in \phi^{-1}[O_2]$. ϕ being a diffeomorphism ensures that M and N have the same dimension, and hence the μ of both $\{x^\mu\}$ and $\{y^\mu\}$ range from 1 to n . A diffeomorphism is originally defined to be a transformation of points; however, it can also be viewed as a transformation of coordinates since we can define a set of new coordinates $\{x'^\mu\}$ on $\phi^{-1}[O_2]$ using $\phi : M \rightarrow N$ as follows: $\forall q \in \phi^{-1}[O_2]$ define $x'^\mu(q) := y^\mu(\phi(q))$. Thus, a diffeomorphism ϕ automatically induces a coordinate transformation $x^\mu \rightarrow x'^\mu$ in the neighborhood $O_1 \cap \phi^{-1}[O_2]$ of p . It is not difficult to prove from Theorem 4.1.2 that $\forall q \in O_1 \cap \phi^{-1}[O_2]$ we have

$$\phi_*[(\partial/\partial x'^\mu)_a|_q] = (\partial/\partial y^\mu)_a|_{\phi(q)}, \quad (4.1.4)$$

from which one can also show that

$$\phi_*[(dx'^\mu)_a|_q] = (dy^\mu)_a|_{\phi(q)}. \quad (4.1.5)$$

Therefore, there exist two different viewpoints for a diffeomorphism $\phi : M \rightarrow N$: ① the **active viewpoint**, which consider ϕ as, by definition, a transformation of points [which turns p into $\phi(p)$] and the consequent tensor transformation [which turns a tensor T at p into a tensor ϕ_*T at $\phi(p)$]; ② the **passive viewpoint**, which regards p and all tensors at p as unchanged, and the consequence of $\phi : M \rightarrow N$ is simply a coordinate transformation (which turns $\{x^\mu\}$ into $\{x'^\mu\}$). Although these two

viewpoints seems quite at odds with each other, they are equivalent for all practical purposes. The theorem below can be seen as some kind of manifestation of this equivalence.

Theorem 4.1.3

$$(\phi_* T)^{\mu_1 \dots \mu_k}_{v_1 \dots v_l} |_{\phi(p)} = T'^{\mu_1 \dots \mu_k}_{v_1 \dots v_l} |_p, \quad \forall T \in \mathcal{F}_M(k, l), \quad (4.1.6)$$

where the left-hand side are the components of the new tensor $\phi_* T$ at the new point $\phi(p)$ in the old coordinate system $\{y^\mu\}$, and the right-hand side are the components of the old tensor T at the old point p in the new coordinate system $\{x'^\mu\}$.

Proof Exercise 4.2. □

Remark 2 Equation (4.1.6) is an equality of real numbers, the left-hand side is the number coming from the active viewpoint (which regards the point and the tensor as changed but the coordinate system as unchanged), while the right-hand side is the number coming from the passive viewpoint (which regards the point and the tensor as unchanged but the coordinate system as changed). Both sides being equal indicates that these two viewpoints are equivalent for all practical purposes.

Example 1 Suppose $T^{a_1 \dots a_k}_{b_1 \dots b_l}$ in Theorem 4.1.3 is a vector v^a . Let $u^a \equiv \phi_* v^a \in V_{\phi(p)}$, then it is not difficult to prove from (4.1.6) that

$$u^\mu = v^\nu (\partial x'^\mu / \partial x^\nu)|_p. \quad (4.1.7)$$

[Optional Reading 4.1.1]

Now we further explain the equivalence of the active and passive viewpoints. Suppose T_{ab} is a tensor field on M , then its components in a coordinate system $\{x^\sigma\}$ form a set of functions of coordinates x^σ , i.e., $T_{\mu\nu}(x^\sigma)$. Suppose there is a coordinate transformation $\{x^\sigma\} \rightarrow \{x'^\sigma\}$, then the components of T_{ab} in the coordinate system $\{x'^\sigma\}$ form a set of functions of coordinates x'^σ , i.e., $T'_{\mu\nu}(x'^\sigma)$. These two sets of functions are different in general. (Here we mean the expressions for $T_{\mu\nu}$ and $T'_{\mu\nu}$ are different, though the symbols for the argument do not matter.) If we want to obtain another set of functions $T'_{\mu\nu}$ from the function set $T_{\mu\nu}$, we just need to perform the coordinate transformation, but not the transformation for points and tensors on the manifold; that is, there is no need to employ the map between manifolds and the map of tensors induced by it. This can be called the “passive approach” of acquiring a new set of functions $T'_{\mu\nu}$. However, the same effect can also be obtained by adopting the following “active approach”. Suppose N is another manifold and there exists a diffeomorphism $\phi : M \rightarrow N$, then $\tilde{T}_{ab} \equiv \phi_* T_{ab}$ is a tensor field on N , the components of which in a coordinate system $\{y^\sigma\}$ are also a set of functions $\tilde{T}_{\mu\nu}(y^\sigma)$ that, in general, have a different form than $T_{\mu\nu}(x^\sigma)$. This approach involves the transformation of points ($\phi : M \rightarrow N$) and the transformation of tensor fields ($\phi_* : T_{ab} \mapsto \tilde{T}_{ab}$) but not any coordinate transformation, which is exactly what the active viewpoint means. In order to make sure that they will lead to the same end—that is, that the new function sets $\tilde{T}_{\mu\nu}$ and $T'_{\mu\nu}$ coming from the active and passive approaches are the same—we only need to set the coordinate transformation on M induced by the diffeomorphism $\phi : M \rightarrow N$ in the active approach as the coordinate transformation $\{x^\sigma\} \rightarrow \{x'^\sigma\}$ in the passive approach. In fact, if we suppose $p \in M$ and $q \equiv \phi(p) \in N$, then

$$\tilde{T}_{\mu\nu}(y^\sigma(q)) = \tilde{T}_{\mu\nu}|_q = (\phi_*T)_{\mu\nu}|_q = T'_{\mu\nu}|_p = T'_{\mu\nu}(x'^\sigma(p)) = T'_{\mu\nu}(y^\sigma(q)),$$

where Theorem 4.1.3 and the requirement of “setting the coordinate transformation induced by $\phi : M \rightarrow N$ as $\{x^\sigma\} \rightarrow \{x'^\sigma\}$ ” are applied in the third and the fifth equality, respectively. This equation above indicates that $\tilde{T}_{\mu\nu}(y^\sigma) = T'_{\mu\nu}(y^\sigma)$, i.e., functions $\tilde{T}_{\mu\nu}$ and $T'_{\mu\nu}$ are equivalent.

This is only an example that shows the equivalence of the active and passive viewpoints for practical purposes. The fact that Theorem 4.1.3 was used in a key step of the proof indicates once again that this theorem is some kind of manifestation of this equivalence.

[The End of Optional Reading 4.1.1]

[Optional Reading 4.1.2]

In this optional reading, we introduce several useful theorems as supplements.

Theorem 4.1.4 Suppose $\phi : M \rightarrow N$ is a smooth map, then $\forall T \in \mathcal{F}_N(0, l), T' \in \mathcal{F}_N(0, l')$ we have

$$\phi^*(T \otimes T') = \phi^*(T) \otimes \phi^*(T'). \quad (4.1.8)$$

Proof The reader should add abstract indices to the equation and carry out the proof. \square

Theorem 4.1.5 Suppose $\phi : M \rightarrow N$ is a smooth map, then $\forall T \in \mathcal{F}_{V_p}(k, 0), T' \in \mathcal{F}_{V_p}(k', 0)$ we have

$$\phi_*(T \otimes T') = \phi_*(T) \otimes \phi_*(T'). \quad (4.1.9)$$

Proof The reader should add abstract indices to the equation and carry out the proof. \square

Theorem 4.1.6 Suppose $\phi : M \rightarrow N$ is a diffeomorphism, then $\forall T \in \mathcal{F}_M(k, l), T' \in \mathcal{F}_M(k', l')$ we have

$$\phi_*(T \otimes T') = \phi_*(T) \otimes \phi_*(T'). \quad (4.1.10)$$

Remark 3 ① The above equation is an equality of tensor fields on N , while (4.1.9) is just an equality of tensors at a point $\phi(p) \in N$. ② The above equation will still hold if we substitute ϕ^* for ϕ_* ; however, T and T' in this case should be tensor fields on N , and the new equation should be viewed as an equality of tensor fields on M .

Proof Exercise. \square

Theorem 4.1.7 Suppose $\phi : M \rightarrow N$ is a diffeomorphism, then ϕ_* (and ϕ^*) commute with any contraction.

Proof To show that $\phi_*(CT) = C(\phi_*T)$, first we take a tensor field $T^a{}_b$ on M as an example. In this case $\phi_*(CT) = C(\phi_*T)$ is an equality of scalar fields on N , and all we have to do is to show that it holds for the image point $\phi(p) \in N$ of any $p \in M$. Suppose $\{(e_\mu)^a\}$ and $\{(e^\mu)_a\}$ are a basis and its dual basis at p , then $T^a{}_b = T^\mu{}_v (e_\mu)^a (e^v)_b$. It follows from (4.1.10) that

$$\phi_*T^a{}_b = (\phi_*T^\mu{}_v)[\phi_*(e_\mu)^a][\phi_*(e^v)_b],$$

and hence

$$C(\phi_*T) = (\phi_*T^\mu{}_v)[\phi_*(e_\mu)^a][\phi_*(e^v)_a].$$

Taking $(\partial/\partial x'^\mu)^a$ from (4.1.4) and $(dx'^\mu)_a$ from (4.1.5) as $(e_\mu)^a$ and $(e^\mu)_a$, respectively, yields

$$[\phi_*(e_\mu)^a][\phi_*(e^v)_a] = (\partial/\partial y^\mu)^a (dy^v)_a = \delta^v{}_\mu.$$

(Actually it can be proved that the above equation holds for any $\{(e_\mu)^a\}$ and $\{(e^\mu)_a\}$ at p . Therefore,

$$C(\phi_* T) = (\phi_* T^\mu{}_v) \delta^v{}_\mu = \phi_*(T^\mu{}_v \delta^v{}_\mu) = \phi_*(T^\mu{}_\mu) = \phi_*(CT).$$

The reader may generalize this proof to a tensor field of arbitrary type on M . \square

[The End of Optional Reading 4.1.2]

4.2 Lie Derivatives

As we have discussed at the end of Sect. 2.2, a smooth vector field v^a on M gives rise to a one-parameter group of diffeomorphisms ϕ .¹ Suppose $T^{\cdots \cdots}$ is a smooth tensor field on M , then $\phi_t^* T^{\cdots \cdots}$ is also a smooth tensor field of the same type, where ϕ_t is a group element from the one-parameter group of diffeomorphisms ϕ . The difference of these two tensor fields at $p \in M$, namely, $\phi_t^* T^{\cdots \cdots}|_p - T^{\cdots \cdots}|_p$, is a tensor at p , and the quotient $(\phi_t^* T^{\cdots \cdots}|_p - T^{\cdots \cdots}|_p)/t$ in the limit of t approaches zero can be viewed as some kind of derivative of the tensor field $T^{\cdots \cdots}$ at p . Therefore, we have the following definition:

Definition 1

$$\mathcal{L}_v T^{a_1 \cdots a_k}{}_{b_1 \cdots b_l} := \lim_{t \rightarrow 0} \frac{1}{t} (\phi_t^* T^{a_1 \cdots a_k}{}_{b_1 \cdots b_l} - T^{a_1 \cdots a_k}{}_{b_1 \cdots b_l}) \quad (4.2.1)$$

is called the **Lie derivative** of a tensor field $T^{a_1 \cdots a_k}{}_{b_1 \cdots b_l}$ along a vector field v^a . (To avoid confusion, the v in \mathcal{L}_v is not written as v^a .)

Remark 1 Since ϕ_t^* is a linear map, the Lie derivative is a linear map from $\mathcal{F}_M(k, l)$ to $\mathcal{F}_M(k, l)$. From (4.2.1) and Theorem 4.1.7 we can also see that \mathcal{L}_v commutes with contractions.

Theorem 4.2.1

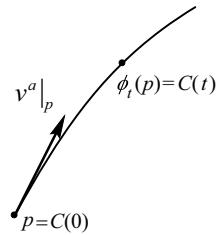
$$\mathcal{L}_v f = v(f), \quad \forall f \in \mathcal{F}. \quad (4.2.2)$$

Proof $\forall p \in M$, suppose $C(t)$ is the orbit of ϕ that passes through p . Set $p = C(0)$, then $\phi_t(p) = C(t)$, and $v^a|_p \equiv (\partial/\partial t)^a|_p$ is the tangent vector of $C(t)$ at p (see Fig. 4.2). Hence,

$$\begin{aligned} \mathcal{L}_v f|_p &= \lim_{t \rightarrow 0} \frac{1}{t} (\phi_t^* f - f)|_p = \lim_{t \rightarrow 0} \frac{1}{t} [f(\phi_t(p)) - f(p)] \\ &= \lim_{t \rightarrow 0} \frac{1}{t} [f(C(t)) - f(C(0))] = \frac{d}{dt} (f \circ C)|_{t=0} = v(f)|_p. \end{aligned} \quad \square$$

¹ If v^a is incomplete, then it can only give rise to a one-parameter local group of diffeomorphisms. This section only involves local properties, so there is no need to distinguish local and global.

Fig. 4.2 Figure for the proof of Theorem 4.2.1



Taking $n = 2$ for example, we now introduce a special coordinate system that is quite useful for computing Lie derivatives. Suppose $\{x^1, x^2\}$ is a coordinate system, then the x^1 -coordinate lines and x^2 -coordinate lines comprise a “coordinate grid”. To see the coordinates of a point in the coordinate patch, all we have to do is to find the intersection of two coordinate lines at which this point is located. Since any Lie derivative is taken along a vector field v^a , we may choose the integral curves of v^a as the x^1 -coordinate lines. More precisely, if $v^a = (\partial/\partial t)^a$, we simply choose t to be the first coordinate x^1 of this system, i.e., we set $v^a = (\partial/\partial x^1)^a$. Then, we arbitrarily choose another set of curves that are transverse to them (that is, the two tangent vectors at any point of intersection are not parallel) as the x^2 -coordinate lines. Such a coordinate system is called a **coordinate system adapted to the vector field v^a** .² The discussion above can be generalized to manifolds of arbitrary dimensions.

Theorem 4.2.2 *The components of the Lie derivative of a tensor field $T^{a_1 \dots a_k}{}_{b_1 \dots b_l}$ along v^a in a coordinate system adapted to v^a are*

$$(\mathcal{L}_v T)^{\mu_1 \dots \mu_k}{}_{\nu_1 \dots \nu_l} = \frac{\partial T^{\mu_1 \dots \mu_k}{}_{\nu_1 \dots \nu_l}}{\partial x^1}. \quad (4.2.3)$$

Remark 2 The left-hand side of the above equation satisfies the tensor transformation law under a coordinate transformation while the right-hand side does not. Hence, this equation cannot be written as an equality of tensors.

Proof Here, we only take $n = 2$, $k = l = 1$ as an example (it is easily carried over to the general case). Since $\phi_t^* = (\phi_t^{-1})_* = \phi_{-t*}$, the components of (4.2.1) in an arbitrary coordinate system can be expressed as

$$(\mathcal{L}_v T)^{\mu}{}_{\nu}|_p = \lim_{t \rightarrow 0} \frac{1}{t} [(\phi_{-t*} T)^{\mu}{}_{\nu}|_p - T^{\mu}{}_{\nu}|_p] \quad \forall p \in M. \quad (4.2.4)$$

Let $q \equiv \phi_t(p)$. Since (4.2.4) only involves the points near p , one can always consider p and q as being in the same adapted coordinate patch. For ϕ_{-t} , q is the old point and p is the new point, and hence it follows from (4.1.6) that

² As long as $v^a \neq 0$ at a point, one can always define a coordinate system adapted to v^a in a neighborhood of the point.

$$(\phi_{-t*}T)^\mu{}_\nu|_p = T'^\mu{}_\nu|_q = \left[\frac{\partial x'^\mu}{\partial x^\rho} \frac{\partial x^\sigma}{\partial x'^\nu} T^\rho{}_\sigma \right]_q, \quad (4.2.5)$$

where x^σ are the adapted coordinates (the old coordinates), while x'^μ are the new coordinates induced by ϕ_{-t} . The right-hand side of the above equation involves the value of the partial derivatives between the new and old coordinates at q which, to calculate, we need to find the coordinate transformation in a small neighborhood N of q . $\forall \bar{q} \in N$, denote $\bar{p} \equiv \phi_{-t}(\bar{q})$. From the definition of adapted coordinates we know that $x^1(\bar{q}) = x^1(\bar{p}) + t$, $x^2(\bar{q}) = x^2(\bar{p})$; also, by definition, the new coordinates at \bar{q} induced by ϕ_{-t} are $x'^1(\bar{q}) \equiv x^1(\bar{p})$, $x'^2(\bar{q}) \equiv x^2(\bar{p})$, and hence $x'^1(\bar{q}) = x^1(\bar{q}) - t$, $x'^2(\bar{q}) = x^2(\bar{q})$. Since \bar{q} is an arbitrary point in N , for N we have $x'^1 = x^1 - t$, $x'^2 = x^2$, and taking the derivatives we get $(\partial x'^\mu / \partial x^\rho)|_q = \delta^\mu{}_\rho$, $(\partial x^\sigma / \partial x'^\nu)|_q = \delta^\sigma{}_\nu$. Therefore, (4.2.5) becomes $(\phi_{-t*}T)^\mu{}_\nu|_p = T'^\mu{}_\nu|_q$, and plugging this into (4.2.4) yields $(\mathcal{L}_v T)^\mu{}_\nu|_p = \partial T^\mu{}_\nu / \partial x^1|_p$. \square

It follows from Theorem 4.2.2 that \mathcal{L}_v satisfies the Leibniz rule.

Theorem 4.2.3

$$\mathcal{L}_v u^a = [v, u]^a, \quad \forall u^a, v^a \in \mathcal{F}(1, 0), \quad (4.2.6)$$

or, by means of the expression for a commutator (3.1.13), we have

$$\mathcal{L}_v u^a = v^b \nabla_b u^a - u^b \nabla_b v^a, \quad (4.2.6')$$

where ∇_a is an arbitrary torsion-free derivative operator.

Proof The claim we are about to prove is an equality of vectors, all we have to show is that the corresponding equality of components in a coordinate system holds. The most convenient one to use is certainly the adapted coordinate system. Suppose the ordinary derivative operator of a coordinate system $\{x^\mu\}$ adapted to v^a is ∂_a , then

$$\begin{aligned} [v, u]^\mu &= (\text{d}x^\mu)_a [v, u]^a = (\text{d}x^\mu)_a (v^b \partial_b u^a - u^b \partial_b v^a) = v^b \partial_b u^\mu \\ &= v(u^\mu) = \partial u^\mu / \partial x^1 = (\mathcal{L}_v u)^\mu, \end{aligned}$$

where the third equality comes from the fact that $v^a = (\partial / \partial x^1)^a$ leads to $\partial_b v^a = 0$, condition (d) in the definition of a derivative operator is used in the fourth equality, and (4.2.3) is used in the last step. \square

Theorem 4.2.4

$$\mathcal{L}_v \omega_a = v^b \nabla_b \omega_a + \omega_b \nabla_a v^b, \quad \forall v^a \in \mathcal{F}(1, 0), \quad \omega_a \in \mathcal{F}(0, 1), \quad (4.2.7)$$

where ∇_a is an arbitrary torsion-free derivative operator.

Proof Exercise 4.7. Hint: use Theorem 4.2.3 and 4.2.1, the latter of which will give $\mathcal{L}_v(\omega_a u^a) = v^b \nabla_b(\omega_a u^a)$. \square

Theorem 4.2.5

$$\mathcal{L}_v T^{a_1 \cdots a_k}{}_{b_1 \cdots b_l} = v^c \nabla_c T^{a_1 \cdots a_k}{}_{b_1 \cdots b_l} - \sum_{i=1}^k T^{a_1 \cdots c \cdots a_k}{}_{b_1 \cdots b_l} \nabla_c v^{a_i} + \sum_{j=1}^l T^{a_1 \cdots a_k}{}_{b_1 \cdots c \cdots b_l} \nabla_{b_j} v^c$$

$\forall T \in \mathcal{F}(k, l), v \in \mathcal{F}(1, 0), \quad \nabla_a \text{ is an arbitrary torsion-free derivative operator.} \quad (4.2.8)$

Proof Exercise. □

4.3 Killing Vector Fields

Up to this point, this chapter has not yet mentioned any metric or any derivative operator associated with a metric since the definition of a Lie derivative does not require any additional structure on the manifold M . However, if a metric field g_{ab} is assigned to M , then one can also impose a higher requirement on a diffeomorphism $\phi : M \rightarrow M$, i.e., $\phi^* g_{ab} = g_{ab}$. Therefore, we have the following definition:

Definition 1 A diffeomorphism $\phi : M \rightarrow M$ is called an **isometric isomorphism**, or **isometry** for short, if $\phi^* g_{ab} = g_{ab}$.

Remark 1 ① An isometry is a special diffeomorphism that “preserves the metric”, namely $\phi^* g_{ab} = g_{ab}$. Note that this is an equality of tensor fields, which means the two tensors $g_{ab}|_p$ and $\phi^* g_{ab}|_p$ at each point p are equal. ② From $\phi^{-1*} \circ \phi^* = (\phi \circ \phi^{-1})^* = \text{identity}$ (see Exercise 4.5(c)) it is easy to see that $\phi : M \rightarrow M$ is an isometry if and only if $\phi^{-1} : M \rightarrow M$ is an isometry.

Among all the vector fields on a manifold M there is a special class of vector fields, namely the smooth vector fields. Each smooth vector field gives rise to a one-parameter group of diffeomorphisms.³ If a metric field g_{ab} is assigned to M , then we can also pick a special subclass among all the smooth vector fields, in which the one-parameter group of diffeomorphisms given by each vector field is a one-parameter group of isometries; that is, each group element $\phi_t : M \rightarrow M$ is an isometry. Therefore, we have the following definition:

Definition 2 A vector field ξ^a on (M, g_{ab}) is called a **Killing vector field** if its one-parameter (local) group of diffeomorphisms is a one-parameter (local) group of isometries. Equivalently (motivated readers should verify this), ξ^a is called a Killing vector field if $\mathcal{L}_\xi g_{ab} = 0$.

³ We do not require the vector field to be complete. When talking about an incomplete vector field, the one-parameter group of diffeomorphisms refers to its one-parameter local group of diffeomorphisms.

Theorem 4.3.1 *The necessary and sufficient condition for ξ^a to be a Killing vector field on (M, g_{ab}) is that ξ^a satisfies the following **Killing equation**:*

$$\nabla_a \xi_b + \nabla_b \xi_a = 0, \quad \text{or} \quad \nabla_{(a} \xi_{b)} = 0, \quad \text{or} \quad \nabla_a \xi_b = \nabla_{[a} \xi_{b]}, \quad (4.3.1)$$

where ∇_a is the torsion-free operator associated with g_{bc} ($\nabla_a g_{bc} = 0$).

Proof For any vector field ξ^a , it follows from (4.2.8) that

$$\mathcal{L}_\xi g_{ab} = \nabla_a \xi_b + \nabla_b \xi_a, \quad (4.3.1')$$

where we used the fact that $\nabla_a g_{bc} = 0$. By definition, ξ^a being a Killing vector field is equivalent to $\mathcal{L}_\xi g_{ab} = 0$. Hence, the necessary and sufficient condition for ξ^a to be a Killing vector field is that it satisfies the Killing equation $\nabla_a \xi_b + \nabla_b \xi_a = 0$. \square

Theorem 4.3.2 *If there exists a coordinate system $\{x^\mu\}$ such that all the components of g_{ab} satisfy $\partial g_{\mu\nu}/\partial x^1 = 0$, then $(\partial/\partial x^1)^a$ is a Killing vector on the coordinate patch.*

Proof $\{x^\mu\}$ is a coordinate system adapted to $(\partial/\partial x^1)^a$. From (4.2.3) we can see that $(\mathcal{L}_{\partial/\partial x^1} g)_{\mu\nu} = \partial g_{\mu\nu}/\partial x^1 = 0$, and hence $\mathcal{L}_{\partial/\partial x^1} g_{ab} = 0$, i.e., $(\partial/\partial x^1)^a$ is a Killing vector field. \square

Theorem 4.3.3 *Suppose ξ^a is a Killing vector field, and T^a is the tangent of a geodesic, then $T^a \nabla_a (T^b \xi_b) = 0$, i.e., $T^b \xi_b$ is a constant along the geodesic.*

Proof $T^a \nabla_a (T^b \xi_b) = \xi_b T^a \nabla_a T^b + T^b T^a \nabla_a \xi_b = T^b T^a \nabla_a \xi_b = 0$, where the definition of a geodesic is used in the second equality, and Theorem 4.3.1 (i.e., $\nabla_a \xi_b = \nabla_{[a} \xi_{b]}$) and Theorem 2.6.2 (d) are used in the third equality. \square

Suppose ξ^a and η^a are Killing vector fields, α and β are real constants, then from the linearity of the Killing equation we know that $\alpha \xi^a + \beta \eta^a$ is also a Killing vector field. It is not difficult to see that the collection of all the Killing vector fields on M is a vector space. It can also be proved (Exercise 4.13) that the commutator $[\xi, \eta]^a$ is also a Killing vector field.

Theorem 4.3.4 *There are at most $n(n+1)/2$ independent Killing vector fields ($n \equiv \dim M$) on (M, g_{ab}) . That is, the dimension of the collection of all the Killing vector fields on M (as a vector space) is less than or equal to $n(n+1)/2$.*

Proof See Wald (1984) pp. 442–443. \square

Remark 2 ① Isometries can be viewed as some kind of symmetry transformations that “preserve the metric”, and thus a Killing vector field represents a symmetry of (M, g_{ab}) . A generalized Riemannian space that has $n(n+1)/2$ independent Killing vector fields is called a maximally symmetric space. ② The general method of finding all the Killing vector fields on (M, g_{ab}) is to find the general solution of the Killing equation. However, for some (M, g_{ab}) that are relatively simple, there also exist methods that are a lot easier. We provide several examples below.

Example 1 Find all the independent Killing vector fields of the following generalized Riemannian spaces.

(1) 2-dimensional Euclidean space $(\mathbb{R}^2, \delta_{ab})$. Suppose $\{x, y\}$ is a Cartesian coordinate system, then $ds^2 = dx^2 + dy^2$, i.e., all the components of the Euclidean metric δ_{ab} in this coordinate system are constant. Hence, it follows from Theorem 4.3.2 that $(\partial/\partial x)^a$ and $(\partial/\partial y)^a$ are Killing vector fields. We believe that a Euclidean space is maximally symmetric, and it follows from Theorem 4.3.4 that there should be three independent Killing Fields when $n = 2$. As expected, if we change to a polar coordinate system, then $ds^2 = dr^2 + r^2 d\varphi^2$, and thus all of the components of δ_{ab} in this coordinate system are independent of φ . Therefore, it follows from Theorem 4.3.2 that $(\partial/\partial \varphi)^a$ is a Killing vector field. The expanded form of it in the coordinate basis of a Cartesian coordinate basis is $(\partial/\partial \varphi)^a = -y(\partial/\partial x)^a + x(\partial/\partial y)^a$. The coefficients of the expansion depends on the coordinates, from which it is not difficult to show that $(\partial/\partial \varphi)^a$ is independent of the first two Killing fields. $(\partial/\partial x)^a$ and $(\partial/\partial y)^a$ being Killing reflects the translational invariance of the 2-dimensional Euclidean metric along the x - and y -axes, while $(\partial/\partial \varphi)^a$ being Killing manifests the rotational invariance of this metric.

(2) 3-dimensional Euclidean space $(\mathbb{R}^3, \delta_{ab})$. Since $n = 3$, there are six independent Killing fields, namely $(\partial/\partial x)^a$, $(\partial/\partial y)^a$, $(\partial/\partial z)^a$, $-y(\partial/\partial x)^a + x(\partial/\partial y)^a$, $-z(\partial/\partial y)^a + y(\partial/\partial z)^a$ and $-x(\partial/\partial z)^a + z(\partial/\partial x)^a$. The first three reflect the translational invariance of the 3-dimensional Euclidean metric along the x -, y - and z -axes, while the last three reflect the rotational invariance of the metric along the z , x , y axes, respectively.

(3) 2-dimensional Minkowski space $(\mathbb{R}^2, \eta_{ab})$. In a Lorentzian coordinate system $\{t, x\}$ we have $ds^2 = -dt^2 + dx^2$, and thus we see that $(\partial/\partial t)^a$ and $(\partial/\partial x)^a$ are Killing fields. To find the third one, we define new coordinates ψ and η as follows:

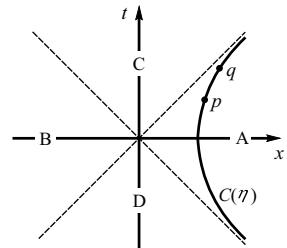
$$x = \psi \cosh \eta, \quad t = \psi \sinh \eta, \quad 0 < \psi < \infty, \quad -\infty < \eta < \infty. \quad (4.3.2)$$

The Minkowski line element can be expressed in terms of the new coordinates as $ds^2 = d\psi^2 - \psi^2 d\eta^2$. This expression indicates that all the components of η_{ab} in the new coordinate system are independent of the coordinate η , and hence $(\partial/\partial \eta)^a$ is also a Killing vector field (whose integral curves are hyperbolas). The expanded form of it in the coordinate basis of a Lorentzian coordinate basis is

$$(\partial/\partial \eta)^a = t(\partial/\partial x)^a + x(\partial/\partial t)^a. \quad (4.3.3)$$

From the fact that the coefficients of the expansion are coordinate dependent we can see that $(\partial/\partial \eta)^a$ is independent of the first two Killing fields. The coordinate patch of η and ψ defined by (4.3.2) is just an open subset of \mathbb{R}^2 which is restricted by $x > |t|$ (see region A in Fig. 4.3). However, (4.3.3) is defined on the whole \mathbb{R}^2 , and it is not difficult to verify that $(\partial/\partial \eta)^a$ is a Killing field on \mathbb{R}^2 . It is timelike in the regions A and B in Fig. 4.3, spacelike in the regions C and D, and null on the two lines with 45° tilt. $t(\partial/\partial x)^a + x(\partial/\partial t)^a$ is called the **boost** Killing vector

Fig. 4.3 The boost Killing vector field $(\partial/\partial\eta)^a$ is timelike in the regions A and B, spacelike in the regions C and D, and null on the two 45° lines



field, which indicates that the Minkowski metric has the invariance under a boost, corresponding to the Lorentz transformation (for details, see Theorem 4.3.5).

(4) 4-dimensional Minkowski space $(\mathbb{R}^4, \eta_{ab})$. Since $n = 4$, there are in total 10 independent Killing fields, divided into three groups:

- (a) 4 translations $(\partial/\partial t)^a, (\partial/\partial x)^a, (\partial/\partial y)^a, (\partial/\partial z)^a$;
- (b) 3 spatial rotations
 $-y(\partial/\partial x)^a + x(\partial/\partial y)^a, -z(\partial/\partial y)^a + y(\partial/\partial z)^a, -x(\partial/\partial z)^a + z(\partial/\partial x)^a$;
- (c) 3 boosts $t(\partial/\partial x)^a + x(\partial/\partial t)^a, t(\partial/\partial y)^a + y(\partial/\partial t)^a, t(\partial/\partial z)^a + z(\partial/\partial t)^a$.

Group (a) reflects the translational invariance of the Minkowski metric along the t -, x -, y -, z -axes; group (b) reflects the spatial rotational invariance with respect to z -, x -, y -axes, respectively; group (c) reflects the invariance under the boosts within the tx -, ty -, tz -planes.

In Sect. 4.1 we already introduced the active and passive viewpoints of a diffeomorphism (in which the former is a transformation of points and tensor fields while in the latter is a coordinate transformation) and their relationship (a transformation of points induces a coordinate transformation). Now that we know an isometry is a special diffeomorphism, we can expect that the coordinate transformation induced by an isometry is also a special coordinate transformation. In fact, this is true! First, we will use the 2-dimensional Euclidean space $(\mathbb{R}^2, \delta_{ab})$ as an example. Each Killing vector field will give rise to a one-parameter group of isometries $\{\phi_\lambda : \mathbb{R}^2 \rightarrow \mathbb{R}^2 \mid \lambda \in \mathbb{R}\}$. From the active viewpoint, there are three kinds of isometries in this group, i.e., three independent Killing vector fields:

① Translational Killing vector field $(\partial/\partial x)^a$. It induces the translation along the x -direction, which can be expressed as $x' = x + \lambda, y' = y$; (It is not difficult to prove by following the proof of Theorem 4.3.5; the expressions in ② and ③ can be proved similarly.)

② Translational Killing vector field $(\partial/\partial y)^a$. It induces the translation along the y -direction, which can be expressed as $x' = x, y' = y + \lambda$;

③ Rotational Killing vector field $(\partial/\partial\varphi)^a = -y(\partial/\partial x)^a + x(\partial/\partial y)^a$. It induces the rotation with respect to the origin, which can be expressed using polar coordinates as $r' = r, \varphi' = \varphi + \lambda$, or expressed using Cartesian coordinates as $x' = x \cos \lambda - y \sin \lambda, y' = x \sin \lambda + y \cos \lambda$.

Now we look at the 2-dimensional Minkowski space $(\mathbb{R}^2, \eta_{ab})$. It also contains three kinds of isometries, i.e., three independent Killing vector fields:

① Time translational Killing vector field $(\partial/\partial t)^a$. It induces the time translation along the t -direction, which can be expressed as $t' = t + \lambda$, $x' = x$ (where x and t are Lorentzian coordinates);

② Spatial translational Killing vector field $(\partial/\partial x)^a$. It induces the spatial translation along the x -direction, which can be expressed as $t' = t$, $x' = x + \lambda$;

③ Boost Killing vector field $(\partial/\partial \eta)^a = t(\partial/\partial x)^a + x(\partial/\partial t)^a$. The coordinate transformation it induces is the well-known Lorentz transformation, see the following theorem:

Theorem 4.3.5 Suppose $\{x, t\}$ is the Lorentzian coordinate system of the 2-dimensional Minkowski space $(\mathbb{R}^2, \eta_{ab})$, $\phi_\lambda : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a group element of the one-parameter group of isometries (i.e., ϕ_λ the isometry labeled by the parameter $\lambda \in \mathbb{R}$) that corresponds to the boost Killing field $\xi^a \equiv t(\partial/\partial x)^a + x(\partial/\partial t)^a$, then the coordinate transformation $\{x, t\} \mapsto \{x', t'\}$ induced by ϕ_λ is a Lorentz transformation.

Remark 3 This theorem indicates that a boost and a Lorentz transformation are two different wordings (active and passive) of the same transformation.

Proof The parametric equations of an integral curve of the vector field satisfying $\xi^a \equiv (\partial/\partial \eta)^a$ are $dx^\mu(\eta)/d\eta = \xi^\mu (\mu = 0, 1)$. Noticing that $\xi^a \equiv t(\partial/\partial x)^a + x(\partial/\partial t)^a$ [see (4.3.3)], we have

$$\frac{dx(\eta)}{d\eta} = t(\eta), \quad \frac{dt(\eta)}{d\eta} = x(\eta). \quad (4.3.4)$$

$\forall p \in \mathbb{R}^2$, suppose $C(\eta)$ is the integral curve that satisfies $p = C(0)$, i.e., $x(0) = x_p$, $t(0) = t_p$, then it is not difficult to prove that (4.3.4) have the particular solutions (i.e., the parametric equations of the curve)

$$x(\eta) = x_p \cosh \eta + t_p \sinh \eta, \quad t(\eta) = x_p \sinh \eta + t_p \cosh \eta. \quad (4.3.5)$$

Suppose $q \equiv \phi_\lambda(p)$, then q is the point on $C(\eta)$ that has the parameter $\eta = \lambda$, i.e., $q = C(\lambda)$. Hence, the new coordinates t' and x' induced by ϕ_λ satisfy

$$x'_p \equiv x_q = x_p \cosh \lambda + t_p \sinh \lambda, \quad t'_p \equiv t_q = x_p \sinh \lambda + t_p \cosh \lambda.$$

Since p is arbitrary, we can drop the subscript p and write

$$\begin{aligned} x' &= x \cosh \lambda + t \sinh \lambda = \cosh \lambda(x + t \tanh \lambda), \\ t' &= t \cosh \lambda + x \sinh \lambda = \cosh \lambda(t + x \tanh \lambda). \end{aligned} \quad (4.3.6)$$

Let $v \equiv \tanh \lambda$, $\gamma \equiv (1 - v^2)^{-1/2} = \cosh \lambda$, then

$$x' = \gamma(x + vt), \quad t' = \gamma(t + vx). \quad (4.3.7)$$

This is exactly the well-known Lorentz transformation. (Note that we have applied the system of geometrized units, where the speed of light $c = 1$). \square

[Optional Reading 4.3.1]

For any point p in \mathbb{R}^2 , $C(\eta)$ in the proof above is a complete curve, i.e., $\eta \in (-\infty, \infty)$. If p is in the region A or B, then $C(\eta)$ is timelike; if p is in the region C or D, then $C(\eta)$ is spacelike; if p is on the lines with 45° tilt, then $C(\eta)$ is null. The most special case is that $p = (0, 0)$, i.e., p is the origin of the $\{t, x\}$ system, where $C(\eta) = p$ (a single-point curve). Thus, each line with 45° tilt is not one integral curve but the union of 3 integral curves, in which the first and second ones are the upper and lower halves (excluding the origin), respectively, and the third one is the single-point curve $\{p\}$. The range of the parameter of these 3 lines are all $(-\infty, \infty)$.

[The End of Optional Reading 4.3.1]

It is easy to obtain from $ds^2 = -dt^2 + dx^2$ and (4.3.7) that $ds^2 = -dt'^2 + dx'^2$, and thus the coordinate transformation induced by the isometry that corresponds to a boost turns a Lorentzian system $\{t, x\}$ into another Lorentzian system $\{t', x'\}$. This conclusion can be generalized to the following theorem:

Theorem 4.3.6 Suppose $\{x^\mu\}$ is a Lorentzian coordinate system of $(\mathbb{R}^n, \eta_{ab})$, then the necessary and sufficient condition for $\{x'^\mu\}$ to also be a Lorentzian coordinate system is that it is induced by $\{x^\mu\}$ through an isometry $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$.

Proof Denote η_{ab} by g_{ab} , and denote its components in coordinate systems $\{x^\mu\}$ and $\{x'^\mu\}$ as $g_{\mu\nu}$ and $g'_{\mu\nu}$, respectively.

(A) Suppose $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is an isometry (i.e., $\phi^* g_{ab} = g_{ab}$), and $\{x'^\mu\}$ is the coordinate system induced by the Lorentzian system $\{x^\mu\}$ through ϕ , then $\forall p \in \mathbb{R}^n$ we have $g'_{\mu\nu}|_p = (\phi_* g)_{\mu\nu}|_{\phi(p)} = (\phi^{-1*} g)_{\mu\nu}|_{\phi(p)} = g_{\mu\nu}|_{\phi(p)} = \eta_{\mu\nu}$, where (4.1.6) is used in the first equality, the third equality comes from the fact that ϕ being an isometry makes ϕ^{-1} an isometry, and the fourth equality comes from the fact that $\{x^\mu\}$ is Lorentzian. This equation shows that the components of g_{ab} at p in the system $\{x'^\mu\}$ are $\eta_{\mu\nu}$, and hence $\{x'^\mu\}$ is a Lorentzian system.

(B) Suppose $\{x^\mu\}$ and $\{x'^\mu\}$ are both Lorentzian coordinate systems, $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the diffeomorphism that corresponds to the coordinate transformation $\{x^\mu\} \mapsto \{x'^\mu\}$, then $\forall p \in \mathbb{R}^n$ we have $(\phi^{-1*} g)_{\mu\nu}|_p = (\phi_* g)_{\mu\nu}|_p = g'_{\mu\nu}|_{\phi^{-1}(p)} = \eta_{\mu\nu} = g_{\mu\nu}|_p$, where (4.1.6) is used in the second equality, while the third and fourth equalities come from the fact that $\{x^\mu\}$ and $\{x'^\mu\}$ are Lorentzian. This indicates that $\phi^{-1*} g_{ab} = g_{ab}$, and hence ϕ^{-1} is (which means ϕ is also) an isometry. \square

Remark 4 This theorem can also be applied to Euclidean space, where one only needs to change the Lorentzian system to a Cartesian system. We therefore can say that under an isometry a Lorentzian (or Cartesian) coordinate system remains Lorentzian (or Cartesian).

4.4 Hypersurfaces

Definition 1 Suppose M and S are manifolds, $\dim S \leqslant \dim M \equiv n$. A map $\phi : S \rightarrow M$ is called an **embedding** if ϕ is one-to-one and C^∞ , and $\forall p \in S$, the pushforward map $\phi_* : V_p \rightarrow V_{\phi(p)}$ is non-degenerate [$V_{\phi(p)}$ is the tangent space at the point $\phi(p)$ in M], i.e., $\phi_* v^a = 0 \Rightarrow v^a = 0$.

Remark 1 The above conditions for embedding makes it so the topology and manifold structure of S can be naturally carried to $\phi[S]$, and hence makes $\phi : S \rightarrow \phi[S]$ a diffeomorphism.

Definition 2 An embedding $\phi : S \rightarrow M$ is called an **embedded submanifold** of M , or a **submanifold** of M for short. The image $\phi[S]$ is also often called an embedded submanifold. If $\dim S = n - 1$, then $\phi[S] \subset M$ is called a **hypersurface** of M .

Example 1 Suppose U is an open subset of M , and restrict the manifold structure of M on U , then U is a manifold with the same dimension of M . Consider U as the S in Definition 1, and set $\phi : U \rightarrow M$ to be the identity map, then $U \equiv \phi[U]$ is an embedded submanifold (of the same dimension).

Example 2 Suppose S is the unit sphere S^2 in \mathbb{R}^3 (viewed as M), then the identity map $\phi : S^2 \rightarrow \mathbb{R}^3$ gives rise to an embedded submanifold of \mathbb{R}^3 . Noticing that S^2 has one lower dimension than \mathbb{R}^3 , we conclude that S^2 is a hypersurface of \mathbb{R}^3 .

[Optional Reading 4.4.1]

An embedded submanifold $\phi[S]$ has two topologies, one is the topology that comes naturally from the embedding (see Remark 1), and the other is the topology on $\phi[S]$ (as a subset of M) induced by M (see Example 5 in Sect. 1.2). These two topologies are not necessarily the same. However, if we further require them to be the same, then we impose a stricter requirement on the embedding. An embedding satisfying this additional requirement is called a **regular embedding** [see Chern et al. (1999)]. The term “embedding” in some works [e.g., Hawking and Ellis (1973)] actually refers to a regular embedding. Suppose $S = \mathbb{R}$, and $M = \mathbb{R}^2$, then an embedding $\phi : S \rightarrow M$ is a smooth curve in \mathbb{R}^2 . The one-to-one condition of ϕ in the definition does not allow the embedded submanifold to be a self-intersecting curve (such as the figure-eight shaped curve in Fig. 4.4). Is the curve that is “arbitrarily close to self-intersecting” but not self-intersecting in Fig. 4.5 an embedded submanifold? The answer is: it is an embedded submanifold but not a regular embedded submanifold. From now on, most of the cases in this text where we talk about an embedded submanifold will refer to a regular embedded submanifold.

[The End of Optional Reading 4.4.1]

Suppose $\phi[S]$ is a hypersurface of M , and $q \in \phi[S] \subset M$. As a point in M , q has an n -dimensional tangent space V_q . If $w^a \in V_q$ is a tangent vector of a curve passing through q and lying on $\phi[S]$ (“lying on” means each point of the curve is on $\phi[S]$), then we say w^a is tangent to $\phi[S]$. Use W_q to denote the subset of V_q that formed by all the elements which are tangent to $\phi[S]$. The definition of a hypersurface assures that W_q is an $(n - 1)$ -dimensional submanifold of V_q . Speaking of a hypersurface, one may naturally think of its normal vectors. Suppose $\phi[S]$ is a hypersurface, $q \in \phi[S]$,

Fig. 4.4 A self-intersecting curve is not an embedded submanifold

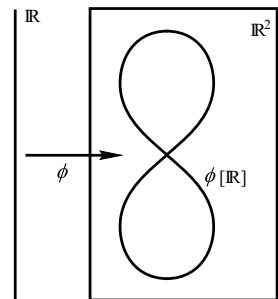
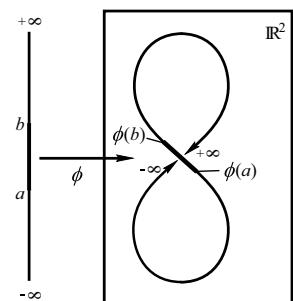


Fig. 4.5 A curve that is “arbitrarily close to self-intersecting” is an embedded submanifold but not a regular embedded submanifold



then a normal vector n^a at q should be defined as a vector that is orthogonal to all the vectors tangent to $\phi[S]$. However, orthogonality is only meaningful after a metric is assigned. When there is no metric on M , one cannot define a normal vector n^a , but can instead define a “normal covector” n_a . **Covector** is another name for dual vector. Since a dual vector gives a real number when acting on a vector (with no need for a metric), a normal covector can be defined as follows:

Definition 3 Suppose $\phi[S]$ is a hypersurface, $q \in \phi[S]$. A nonzero dual vector $n_a \in V_q^*$ is called a **normal covector** of $\phi[S]$ at q if $n_a w^a = 0, \forall w^a \in W_q$.

Theorem 4.4.1 *There exists a normal covector n_a at each point q on a hypersurface $\phi[S]$. The normal covector at q is unique up to a numerical factor.*

Proof Suppose $\{(e_2)^a, \dots, (e_n)^a\}$ is an arbitrary basis of W_q . Since $\dim V_q = n$, there must be elements in V_q that are linearly independent of $\{(e_2)^a, \dots, (e_n)^a\}$. Choose any one of such elements and denote it by $(e_1)^a$, then $\{(e_\mu)^a | \mu = 1, \dots, n\}$ is a basis of V_q , whose dual basis is denoted by $\{(e^\mu)_a\}$. Set $n_a = (e^1)_a$, then $n_a (e_\tau)^a = \delta^1_\tau = 0$ ($\tau = 2, \dots, n$). Hence $n_a w^a = 0 \quad \forall w^a \in W_q$, and thus n_a is a normal covector. If there exists m_a that satisfies $m_a (e_\tau)^a = 0$ ($\tau = 2, \dots, n$), then its components in the dual basis $\{(e^\mu)_a\}$ are $m_\tau = m_a (e_\tau)^a = 0$ ($\tau = 2, \dots, n$), and thus $m_a = m_1 (e^1)_a = m_1 n_a$, i.e., m_a and n_a only differ by multiplication by a numerical factor m_1 . \square

Remark 2 A normal covector of an embedded submanifold that is not a hypersurface does not have a uniqueness like this.

[Optional Reading 4.4.2]

Suppose x, y, z are the natural coordinates of \mathbb{R}^3 . Consider a function $f = ax + by + cz$ (where at least one of the constants a, b, c is nonzero), then the points in \mathbb{R}^3 that satisfy $f = 0$ will form a hypersurface (a plane) in \mathbb{R}^3 . If $f = x^2 + y^2 + z^2 - a^2, a \neq 0$, then the equation $f = 0$ represents another hypersurface (a sphere). However, if $f = x^2 + y^2 + z^2$, then only the origin satisfies $f = 0$, and therefore $f = 0$ does not at all represent a hypersphere. The key point is that in this case we have $d f|_{f=0} = 0$. Another extreme example is the case where $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ is defined as $f(p) = 0 \forall p \in \mathbb{R}^3$. In this case the subset of points that satisfy $f = 0$ is \mathbb{R}^3 itself, and thus is not a hypersurface either. The key point is still $d f|_{f=0} = 0$. Generalizing to the cases where f is a smooth function on an arbitrary manifold M , it can be proved that, as long as $d f|_{f=c} \neq 0$ (i.e., $\nabla_a f|_{f=c} \neq 0$), then $f = c$ (constant) gives a hypersurface in M [for details, see Chillingworth (1976) pp. 156–158].

[The End of Optional Reading 4.4.2]

Theorem 4.4.2 *Let $\phi[S]$ represent the hypersurface defined by $f = \text{constant}$. Suppose $q \in \phi[S]$, and $\nabla_a f|_q \neq 0$, then $\nabla_a f|_q$ is a normal covector of $\phi[S]$ at q .*

Proof All we have to prove is that, for any $q \in \phi[S]$, we have $w^a \nabla_a f = 0, \forall w^a \in W_q$. Since w^a is always tangent to a curve $C(t)$ lying on $\phi[S]$ and passing through q , we get $w^a \nabla_a f = \frac{\partial}{\partial t}(f) = 0 \forall w^a \in W_q$, where the last step is because f is a constant on $C(t)$. \square

Suppose n_a is a normal covector of $\phi[S]$ at q . If there is a metric g_{ab} on M , then $n^a \equiv g^{ab} n_b \in V_q$ is orthogonal to all tangent vectors at q on $\phi[S]$ (since $g_{ab} n^a w^b = n_b w^b = 0 \forall w^b \in W_b$), and hence n^a is called a **normal vector** of the hypersurface $\phi[S]$ at q . If g_{ab} is positive definite (e.g., \mathbb{R}^2 embedded into 3-dimensional Euclidean space), n^a certainly does not belong to W_q , i.e., $n^a \in V_q - W_q$; however, if g_{ab} is Lorentzian, then it is possible that n^a belongs to W_q . Now we will discuss the case where g_{ab} is a Lorentzian metric.

Theorem 4.4.3 *Suppose n^a is a normal vector of $\phi[S]$ at q , then a necessary and sufficient condition for $n^a \in W_q$ is $n^a n_a = 0$.*

Proof

(A) Suppose $n^a \in W_q$. Since n_a is a normal covector of $\phi[S]$, regarding the w^a in Definition 3 as the n^a in the present expression $n^a n_a$, we have $n^a n_a = 0$.

(B) From the proof of Theorem 4.4.1 we know that for any normal covector n_a there exists a basis $\{(e_\mu)^a\}$ such that $(e_2)^a, \dots, (e_n)^a \in W_q$ and $n_a = (e^1)_a$; hence, for the first component of n^a in this basis we have $n^1 = n^a (e^1)_a = n^a n_a$. Therefore, $n^a n_a = 0 \Rightarrow n^1 = 0 \Rightarrow n^a = \sum_{\tau=2}^n n^\tau (e_\tau)^a \in W_q$. \square

Example 3 Suppose $S = \mathbb{R}$, $M = \mathbb{R}^2$, the metric on M is $g_{ab} = \eta_{ab}$, and $\phi : \mathbb{R} \rightarrow \mathbb{R}^2$ is an embedding, then $\phi[\mathbb{R}]$ is a hypersurface in the 2-dimensional Minkowski spacetime. Suppose t and x are Lorentzian coordinates. Here we discuss the following three representative cases [the noteworthy one is case (3) where the normal vector is null]:

(1) $\phi[\mathbb{R}]$ is parallel to the x -axis [see Fig. 4.6a]. $\forall q \in \phi[\mathbb{R}]$, let $(e_2)^a = (\partial/\partial x)^a$, and choose $(e_1)^a = \alpha(\partial/\partial t)^a + \beta(\partial/\partial x)^a$, (α, β can be arbitrary real numbers, but $\alpha \neq 0$) then it is not difficult to verify that $(e^1)_a = \alpha^{-1}(\partial t)_a$. From the proof of

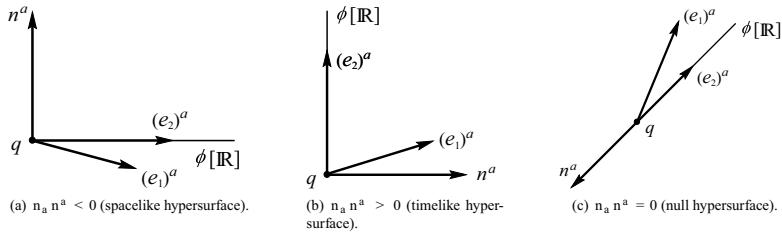


Fig. 4.6 Three cases of embedding \mathbb{R} into \mathbb{R}^2 (t -axis points vertically upwards, x -axis points horizontally to the right)

Theorem 4.4.1 we can see that $(e^1)_a$ is a normal covector n_a whose corresponding normal vector is $n^a = \alpha^{-1} g^{ab} (dt)_b = -\alpha^{-1} (\partial/\partial t)^a$, satisfying $n^a \notin W_q$ and $n_a n^a < 0$ (i.e., n^a is timelike).

(2) $\phi[\mathbb{R}]$ is parallel to the t -axis [see Fig. 4.6b]. $\forall q \in \phi[\mathbb{R}]$, let $(e_2)^a = (\partial/\partial t)^a$, and choose $(e_1)^a = \alpha(\partial/\partial t)^a + \beta(\partial/\partial x)^a$, (α, β can be arbitrary real numbers, but $\beta \neq 0$.) then $(e^1)_a = \beta^{-1}(dx)_a$. Take $(e^1)_a$ to be the normal covector n_a whose corresponding normal vector is $n^a = \beta^{-1}(\partial/\partial x)^a$, satisfying $n^a \notin W_q$ and $n_a n^a > 0$ (i.e., n^a is spacelike).

(3) $\phi[\mathbb{R}]$ makes an angle of 45° with the x -axis (in Euclidean) [see Fig. 4.6c]. $\forall q \in \phi[\mathbb{R}]$, let $(e_2)^a = (\partial/\partial t)^a + (\partial/\partial x)^a$, and choose $(e_1)^a = \alpha(\partial/\partial t)^a + \beta(\partial/\partial x)^a$, $\alpha \neq \beta$, then $(e^1)_a = (\alpha - \beta)^{-1}[(dt)_a - (dx)_a]$. Take $(e^1)_a$ to be the normal covector n_a whose corresponding normal vector is

$$n^a = (\alpha - \beta)^{-1} g^{ab} [(dt)_b - (dx)_b] = -(\alpha - \beta)^{-1} [(\partial/\partial t)^a + (\partial/\partial x)^a] = -(\alpha - \beta)^{-1} (e_2)^a,$$

satisfying $n^a \in W_q$ and $n_a n^a = 0$ (i.e., n^a is null). In this case, the normal vector n^a of the hypersurface is not only perpendicular to all the vectors at q tangent to the surface, but itself is also one of these tangent vectors!

Definition 4 A hypersurface is said to be **spacelike** if its normal vectors are everywhere timelike ($n^a n_a < 0$); a hypersurface is said to be **timelike** if its normal vectors are everywhere spacelike ($n^a n_a > 0$); a hypersurface is said to be **null** or **lightlike** if its normal vectors are everywhere null ($n^a n_a = 0$).

If $n^a n_a \neq 0$, when we talk about a normal vector later on, we will regard it as a normalized normal vector, i.e., $n^a n_a = \pm 1$.

Definition 5 Suppose $\phi[S]$ is an embedding submanifold (not necessarily a hypersurface) in M . Let W_q be the tangent space at an arbitrary point $q \in \phi[S]$ that is tangent to $\phi[S]$. A tensor h_{ab} on W_q is called the **induced metric** derived from the metric g_{ab} on V_q if

$$h_{ab} w_1^a w_2^b = g_{ab} w_1^a w_2^b, \quad \forall w_1^a, w_2^b \in W_q. \quad (4.4.1)$$

The induced metric h_{ab} is essentially the result of restricting the acting target of g_{ab} of V_q to W_q . Since h_{ab} is defined pointwisely on $\phi[S]$, it gives rise to an induced metric field on $\phi[S]$. When $\phi[S]$ is a timelike or spacelike hypersurface, the induced metric can be conveniently expressed by the normalized normal vector $(n^a n_a = \pm 1)$ as

$$h_{ab} = g_{ab} \mp n_a n_b . \quad (- \text{ when } n^a n_a = +1, \text{ and } + \text{ when } n^a n_a = -1.) \quad (4.4.2)$$

It is easy to see that $\forall w_1^a, w_2^b \in W_q$ we have $h_{ab} w_1^a w_2^b = g_{ab} w_1^a w_2^b \mp n^a w_1^a n_b w_2^b = g_{ab} w_1^a w_2^a$, which satisfies (4.4.1). However, there are actually many h_{ab} that satisfy (4.4.1), why do we only use the one defined by (4.4.2)? For the reason, see Optional Reading 4.4.3.

[Optional Reading 4.4.3]

For convenience, we suppose V_q to be 4-dimensional (and thus W_q is 3-dimensional). As an induced metric (a metric on W_q), h_{ab} in (4.4.1) is a tensor on W_q (a 3-dimensional tensor), i.e., $h_{ab} \in \mathcal{T}_{W_q}(0, 2)$ (which cannot act on elements in $V_q - W_q$). However, for the convenience of performing the 4-dimensional calculation, we want to find a 4-dimensional tensor of type $(0, 2)$ [i.e., an element of $\mathcal{T}_{V_q}(0, 2)$], which can represent the 3-dimensional tensor h_{ab} . $h_{ab} \equiv g_{ab} \mp n_a n_b$ is such a 4-dimensional tensor (note that both terms on the right-hand side are 4-dimensional tensors). To distinguish from the h_{ab} in (4.4.1), we temporarily denote the h_{ab} in $h_{ab} \equiv g_{ab} \mp n_a n_b$ as \bar{h}_{ab} . It can be proved that $\mathcal{T}_{V_q}(0, 2)$ has a subset $\mathcal{S}_{V_q}(0, 2) \equiv \{T_{ab} \in \mathcal{T}_{V_q}(0, 2) \mid T_{ab} n^a = 0, T_{ab} n^b = 0\}$ that is naturally isomorphic to $\mathcal{T}_{W_q}(0, 2)$, and thus $\mathcal{S}_{V_q}(0, 2)$ and $\mathcal{T}_{W_q}(0, 2)$ can be naturally identified (for details see Chap. 14). It is easy to see that $g_{ab} \notin \mathcal{S}_{V_q}(0, 2)$ while $\bar{h}_{ab} \in \mathcal{S}_{V_q}(0, 2)$, and $\bar{h}_{ab} w_1^a w_2^b = g_{ab} w_1^a w_2^b \mp n_1^a n_2^b \in W_q$; thus, one can identify \bar{h}_{ab} as h_{ab} . It can also be proved (left to the reader) that the only element in $\mathcal{S}_{V_q}(0, 2)$ that satisfies (4.1.1) (and thus can serve as h_{ab}) is \bar{h}_{ab} , this is the reason why we regard the 4-dimensional tensor $\bar{h}_{ab} \equiv g_{ab} \mp n_a n_b$ as the induced metric. From now on, we will not distinguish the notation of \bar{h}_{ab} and h_{ab} .

The above conclusion about tensors of type $(0, 2)$ can also be generalized as follows: a special subset of $\mathcal{T}_{V_q}(0, l)$, namely $\{T_{a_1 \dots a_l} \in \mathcal{T}_{V_q}(0, l) \mid \text{the contraction on } n_a \text{ and any index of } T_{a_1 \dots a_l} \text{ vanishes}\}$, is naturally isomorphic to $\mathcal{T}_{W_q}(0, l)$, and thus they can be naturally identified. This identification makes it possible to substitute the elements of the former one for the elements of the latter one when discussing and writing equations, which brings us great convenience.

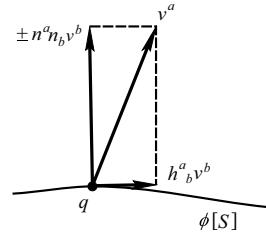
[The End of Optional Reading 4.4.3]

Remark 3 Equation (4.4.2) also holds when g_{ab} is positive definite (just change the sign \mp to $-$). As an exercise, the reader should write down the expression of expanding the 3-dimensional Euclidean metric using the dual vector basis of a spherical coordinate system, and verify that the induced metric $h_{ab} = g_{ab} - n_a n_b$ on the sphere is the same as the induced metric \hat{g}_{ab} defined in Example 2 of Sect. 3.3. [Hint: the normalized normal covector on a sphere is $n_a = (dr)_a$.]

Suppose $\phi[S]$ is a timelike or spacelike hypersurface, $q \in \phi[S]$, and h_{ab} satisfies (4.4.2). Let

$$h^a{}_b \equiv g^{ac} h_{cb} = \delta^a{}_b \mp n^a n_b , \quad (4.4.3)$$

Fig. 4.7 $v^a \in V_q$ is decomposed into the normal component $\pm n^a(n_b v^b)$ and the tangential component $h^a_b v^b \in W_q$



then $\forall v^a \in V_q$ we have $h^a_b v^b = v^a \mp n^a(n_b v^b)$, or

$$v^a = h^a_b v^b \pm n^a(n_b v^b). \quad (4.4.4)$$

The above equation represents a decomposition of the vector v^a (Fig. 4.7), where $\pm n^a(n_b v^b)$ is parallel to n^a , called the normal component, and $h^a_b v^b$ is perpendicular to n^a [since $n_a(h^a_b v^b) = 0$], called the tangential component (the component tangent to $\phi[S]$). h^a_b is called the **projection map** from V_q to W_q .

Theorem 4.4.4 *The induced “metric” on a null hypersurface is degenerate (and thus there is no induced metric).*

Proof Let h_{ab} represent the induced “metric”. The hypersurface being null leads to the result that $n^a \in W_q$ (see Theorem 4.4.3), and hence there is a nonzero element n^a in W_q such that $h_{ab}n^a w^b = g_{ab}n^a w^b = 0, \forall w^a \in W_q$. Thus, h_{ab} is a degenerate tensor on W_q . \square

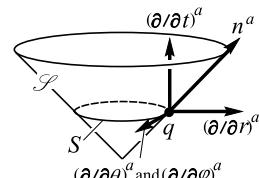
Example 4 Suppose t, x, y, z are Lorentzian coordinates of the 4-dimensional Minkowski space $(\mathbb{R}^4, \eta_{ab})$, r, θ, φ are the spherical coordinates corresponding to x, y, z , then η_{ab} can be expressed in terms of the dual coordinate basis vectors as

$$\eta_{ab} = -(dt)_a(dt)_b + (dr)_a(dr)_b + r^2(d\theta)_a(d\theta)_b + r^2 \sin^2 \theta (d\varphi)_a(d\varphi)_b. \quad (4.4.5)$$

The equation $t - r = 0$ defines a null hypersurface \mathcal{S} , which is a cone with the origin as its apex (see Fig. 4.8). $\forall q \in \mathcal{S} \subset \mathbb{R}^4$ (q is not at the apex), we have a 4-dimensional tangent space V_q and a 3-dimensional tangent space (tangent to \mathcal{S}) $W_q \subset V_q$. Let

$$n^a|_q \equiv (\partial/\partial t)^a|_q + (\partial/\partial r)^a|_q \quad (\text{the subscript } q \text{ will be omitted below}),$$

Fig. 4.8 The induced “metric” of a null hypersurface \mathcal{S} is degenerate



then n^a is a null normal vector of \mathcal{S} at q , and hence $n^a \in W_q$. Therefore, $\{(\partial/\partial\theta)^a, (\partial/\partial\varphi)^a, n^a\}$ is a basis of W_q . Now we calculate the components $h_{\mu\nu}$ of the “metric” h_{ab} induced by η_{ab} on W_q .

$$h_{\theta\theta} \equiv h_{ab}(\partial/\partial\theta)^a(\partial/\partial\theta)^b = \eta_{ab}(\partial/\partial\theta)^a(\partial/\partial\theta)^b = r^2,$$

where (4.4.1) is used in the second equality and (4.4.5) is used in the third equality. Similarly, we have $h_{\phi\phi} = r^2 \sin^2 \theta$, and the third diagonal element of $h_{\mu\nu}$ (denoted by h_{nn}) is

$$h_{nn} \equiv h_{ab}n^a n^b = \eta_{ab}[(\partial/\partial t)^a + (\partial/\partial r)^a][(\partial/\partial t)^b + (\partial/\partial r)^b] = -1 + 1 = 0.$$

Also, it is easy to verify that all of the non-diagonal elements vanish. Hence,

$$(h_{\mu\nu}) = \begin{bmatrix} r^2 & 0 & 0 \\ 0 & r^2 \sin^2 \theta & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and therefore h_{ab} is degenerate [we also say its “signature” is $(+, +, 0)$]. Thus, η_{ab} does not have an induced metric on the null hypersurface \mathcal{S} . However, the intersection S of \mathcal{S} and an arbitrary constant- t surface ($t > 0$) is a 2-dimensional sphere with a radius $r = t$. Let $\hat{W}_q \subset W_q$ represent the subspace formed by all the elements in W_q that are tangent to S (see Fig. 4.8), then η_{ab} does have an induced metric on \hat{W}_q , denoted by \hat{h}_{ab} . Also, it is not difficult to verify that

$$\hat{h}_{ab} = r^2(\mathrm{d}\theta)_a(\mathrm{d}\theta)_b + r^2 \sin^2 \theta (\mathrm{d}\varphi)_a(\mathrm{d}\varphi)_b. \quad (4.4.6)$$

It is not difficult for the reader to discuss the null hypersurface in $(\mathbb{R}^4, \eta_{ab})$ defined by $t - z = 0$ in a similar manner.

In the discussion so far, we have considered an embedded submanifold $\phi[S]$ as the image of the embedding map $\phi : S \rightarrow M$ for convenience. However, sometimes it is useful to regard the map ϕ itself as a submanifold, as it was originally defined in Definition 2. In this case, the induced metric in Definition 5 can be equivalently defined as follows:

Definition 5' Suppose (M, g_{ab}) is a generalized Riemannian space, and $\phi : S \rightarrow M$ is an embedding submanifold of M . Let W_q be the tangent space at an arbitrary point $q \in S$. A metric h_{ab} on W_q is called the **induced metric** derived from g_{ab} if

$$h_{ab}w_1^a w_2^b = g_{ab}(\phi_* w_1^a)(\phi_* w_2^b), \quad \forall w_1^a, w_2^b \in W_q. \quad (4.4.7)$$

Since, by definition, $g_{ab}(\phi_* w_1^a)(\phi_* w_2^b) = (\phi^* g_{ab})w_1^a w_2^b$, and since w_1^a and w_2^a are arbitrary, the above equation can also be written simply as

$$h_{ab} = \phi^* g_{ab}. \quad (4.4.8)$$

Note that the above definition is valid at any point $q \in S$. Since q is arbitrary, the induced metric as a tensor field on S is essentially the pullback of g_{ab} on M .

Theorem 4.4.5 Suppose (M, g_{ab}) is a generalized Riemannian space, and $\phi : S \rightarrow M$ is an embedded submanifold, with $h_{ab} = \phi^* g_{ab}$ the induced metric on S . Suppose $\psi : M \rightarrow M$ is a diffeomorphism, then (1) $\phi' = \psi \circ \phi$ is an embedded submanifold; (2) the induced metric $h'_{ab} = \phi'^* g_{ab}$ is equal to h_{ab} when ψ is an isometry of g_{ab} .

Proof (1) Since both ϕ and ψ are one-to-one and C^∞ , so is $\phi' = \psi \circ \phi$. Let W_p be the tangent space of S at an arbitrary point $p \in S$. For any $w^a \in W_p$, we have $\phi'_* w^a = (\psi \circ \phi)_* w^a = \psi_*(\phi_* w^a)$ (see Exercise 4.5). Since ψ is a diffeomorphism, $\psi_* : V_{\phi(p)} \rightarrow V_{\phi'(p)}$ is an isomorphism (see Exercise 4.4), and thus $\phi'_* w^a = 0$ implies $\phi_* w^a = 0$. Also, since $\phi_* : W_p \rightarrow V_{\phi(p)}$ is nondegenerate, $\phi_* w^a = 0$ implies $w^a = 0$. As a consequence, $\phi'_* : W_p \rightarrow V_{\phi'(p)}$ is nondegenerate at every $p \in S$. Therefore, $\phi' : S \rightarrow M$ is an embedded submanifold.

(2) Using the result of Exercise 4.5 (c), $h'_{ab} = \phi'^* g_{ab} = (\psi \circ \phi)^* g_{ab} = (\phi^* \circ \psi^*) g_{ab} = \phi^* (\psi^* g_{ab})$. When $\psi : M \rightarrow M$ is an isometry of g_{ab} , we have $\psi^* g_{ab} = g_{ab}$, and so $h'_{ab} = \phi^* g_{ab} = h_{ab}$. \square

Exercises

~4.1. Show that $(\phi_* v)^a$ defined by (4.1.2) satisfies the two conditions for a vector in Definition 2 of Sect. 2.2.

~4.2. Prove Theorems 4.1.1, 4.1.2, 4.1.3.

4.3. Suppose $\phi : M \rightarrow N$ is a smooth map, $p \in M$, and y^μ are the coordinates in a neighborhood of $\phi(p)$. Show that

$$(\phi_* v)^a = v(\phi^* y^\mu)(\partial/\partial y^\mu)^a, \quad \forall v^a \in V_p.$$

4.4. Suppose M and N are manifolds, $\phi : M \rightarrow N$ is a diffeomorphism, $p \in M$, and $q \equiv \phi(p)$. Show that the pushforward map $\phi_* : V_p \rightarrow V_q$ is an isomorphism.

4.5. Suppose M, N, Q are manifolds, $\phi : M \rightarrow N$ and $\psi : N \rightarrow Q$ are smooth maps.

(a) Show that $(\psi \circ \phi)^* f = (\phi^* \circ \psi^*) f, \forall f \in \mathcal{F}_Q$.

(b) Show that $(\psi \circ \phi)_* v^a = \psi_*(\phi_* v^a), \forall p \in M, v^a \in V_p$.

(c) Regard both $(\psi \circ \phi)^*$ and $\phi^* \circ \psi^*$ as maps from $\mathcal{F}_Q(0, l)$ to $\mathcal{F}_M(0, l)$.

Show that

$$(\psi \circ \phi)^* = \phi^* \circ \psi^*.$$

4.6. Suppose $\phi : M \rightarrow N$ is a diffeomorphism, v^a and u^a are vector fields on M . Show that $\phi_*([v, u]^a) = [\phi_* v, \phi_* u]^a$, where $[v, u]^a$ represents the commutator.

~4.7. Prove Theorem 4.2.4.

~4.8. Suppose $v^a \in \mathcal{F}_M(1, 0)$, $\omega_a \in \mathcal{F}_M(0, 1)$. Show that for any coordinate system $\{x^\mu\}$ we have

$$(\mathcal{L}_v \omega)_\mu = v^\nu \frac{\partial \omega_\mu}{\partial x^\nu} + \omega_\nu \frac{\partial v^\nu}{\partial x^\mu}. \quad \text{Hint: use (4.2.7) and set the } \nabla_a \text{ to be } \partial_a.$$

~4.9. Suppose $u^a, v^a \in \mathcal{F}_M(1, 0)$, then the following equality holds when both sides acting on a tensor field of any type:

$$[\mathcal{L}_v, \mathcal{L}_u] = \mathcal{L}_{[v,u]} \quad (\text{where } [\mathcal{L}_v, \mathcal{L}_u] \equiv \mathcal{L}_v \mathcal{L}_u - \mathcal{L}_u \mathcal{L}_v).$$

Prove the case where the acting targets are respectively $f \in \mathcal{F}_M$ and $w^a \in \mathcal{F}_M(1, 0)$. Hint: when the acting target is w^a one can use the Jacobi identity (Exercise 2.8).

- 4.10. Suppose F_{ab} is an antisymmetric tensor field on 4-dimensional Minkowski space, whose components in a Lorentzian coordinate system $\{t, x, y, z\}$ are $F_{01} = -F_{13} = x\rho^{-1}$, $F_{02} = -F_{23} = y\rho^{-1}$, $F_{03} = F_{12} = 0$, where $\rho \equiv (x^2 + y^2)^{1/2}$. Show that F_{ab} has rotational symmetry, i.e., $\mathcal{L}_v F_{ab} = 0$, where $v^a = -y(\partial/\partial x)^a + x(\partial/\partial y)^a$.
- 4.11. Suppose ξ^a is a Killing vector field on (M, g_{ab}) , and ∇_a is associated with g_{ab} . Show that $\nabla_a \xi^a = 0$.
- 4.12. Suppose ξ^a is a Killing vector field on (M, g_{ab}) , $\phi : M \rightarrow M$ is an isometry. Show that $\phi_* \xi^a$ is also a Killing vector field on (M, g_{ab}) . Hint: use the conclusion in Exercise 4.5(c).
- 4.13. Suppose ξ^a and η^a are Killing vector fields on (M, g_{ab}) . Show that their commutator $[\xi, \eta]^a$ is also a Killing vector field. NB: This conclusion makes the collection of all Killing vector fields on M not only a vector space, but also a Lie algebra (for details, see Appendix G in Volume II).
- 4.14. Suppose ξ^a is a Killing vector field of a generalized Riemannian space (M, g_{ab}) , and $R_{abc}{}^d$ is the Riemann curvature tensor of g_{ab} .
 - (a) Show that $\nabla_a \nabla_b \xi_c = -R_{bca}{}^d \xi_d$. NB: This equation is significant for proving Theorem 4.3.4. Hint: from the definition of $R_{abc}{}^d$ and the Killing equation (4.3.1) we can see that $\nabla_a \nabla_b \xi_c + \nabla_b \nabla_c \xi_a = R_{abc}{}^d \xi_d$. Refer this to as the first equation. By substituting the indices $a \mapsto b, b \mapsto c, c \mapsto a$ we get the second equation, and by substituting twice we get the third equation. Adding the first equation to the second equation and subtracting the third equation, and using (3.4.7), one can prove the claim.
 - (b) Use the conclusion of (a) to show that $\nabla^a \nabla_a \xi^c = -R_{cd} \xi^d$, where R_{cd} is the Ricci tensor.
- ~4.15. Verify that $(\partial/\partial \eta)^a$ in (4.3.3) indeed satisfies the Killing equation (4.3.1).
- ~4.16. Find the coordinate transformation induced by an arbitrary element ϕ_a from the one-parameter group of isometries generated by $R^a = x(\partial/\partial y)^a - y(\partial/\partial x)^a$ in the 2-dimensional Euclidean space.

- *4.17. Suppose each point of a hypersurface $\phi[S]$ in a spacetime (M, g_{ab}) has a null tangent vector while it does not have any timelike tangent vector (“tangent vector” means the vector is tangent to $\phi[S]$). Show that this is a null hypersurface. Hints: ① show that any vector orthogonal to a timelike vector t^a must be spacelike [choose an orthonormal basis $\{(e_\mu)^a\}$ such that $(e_0)^a = t^a$]; ② show that each point on a timelike hypersurface has a timelike tangent vector; ③ prove the original claim from these two lemmas.

References

- Chern, S. S., Chen, W. and Lam, K. S. (1999), *Lectures on Differential Geometry*, World Scientific Publishing Company, Singapore.
Chillingworth, D. (1976), *Differential Topology with a View to Applications*, Pitman Publishing, London.
Hawking, S. W. and Ellis, G. F. R. (1973), *The Large Scale Structure of Space-Time*, Cambridge University Press, Cambridge.
Wald, R. M. (1984), *General Relativity*, The University of Chicago Press, Chicago.

Chapter 5

Differential Forms and Their Integrals



5.1 Differential Forms

We first introduce “forms” on an n -dimensional vector space V , and then discuss “differential forms” on an n -dimensional manifold M .

Definition 1 $\omega_{a_1 \dots a_l} \in \mathcal{T}_V(0, l)$ is called an **l -form** on V if

$$\omega_{a_1 \dots a_l} = \omega_{[a_1 \dots a_l]}.$$

For convenience in writing, we will sometimes drop the lower indices and write an l -form as ω .

Theorem 5.1.1 (a) $\omega_{a_1 \dots a_l} = \omega_{[a_1 \dots a_l]} \Rightarrow$ for any basis we have $\omega_{\mu_1 \dots \mu_l} = \omega_{[\mu_1 \dots \mu_l]}$;
(b) \exists a basis such that $\omega_{\mu_1 \dots \mu_l} = \omega_{[\mu_1 \dots \mu_l]} \Rightarrow \omega_{a_1 \dots a_l} = \omega_{[a_1 \dots a_l]}.$

Proof Exercise. □

Theorem 5.1.2 Suppose ω is an l -form, then

$$(a) \omega_{a_1 \dots a_l} = \delta_\pi \omega_{a_{\pi(1)} \dots a_{\pi(l)}}. \quad (5.1.1)$$

[See the explanation after (2.6.14) for the meaning of δ_π , $a_{\pi(1)}, \dots, a_{\pi(l)}$.] For example, $\omega_{ab} = -\omega_{ba}$, $\omega_{abc} = -\omega_{acb} = \omega_{cab} = \dots$;

$$(b) \text{For any basis, } \omega_{\mu_1 \dots \mu_l} = \delta_\pi \omega_{\mu_{\pi(1)} \dots \mu_{\pi(l)}}. \quad (5.1.1')$$

Proof (a) See the proof of Theorem 2.6.1 (b);

(b) Exercise. □

It follows from (5.1.1') that any component $\omega_{\mu_1 \dots \mu_l}$ of an l -form with repeated indices must vanish, e.g.,

$$\omega_{112} = \omega_{133} = \omega_{212} = 0.$$

Denote the collection of all the l -forms on V by $\Lambda(l)$. A 1-form is actually a dual vector on V , and hence $\Lambda(1) = V^*$. We stipulate that any real number is called a **0-form** on V , then $\Lambda(0) = \mathbb{R}$. Since an l -form is a tensor of type $(0, l)$, we naturally have $\Lambda(l) \subset \mathcal{T}_V(0, l)$. Moreover, it is easy to show that $\Lambda(l)$ is a linear subspace of $\mathcal{T}_V(0, l)$. The computation of the dimension of $\Lambda(l)$ can be inspired by the computation of the dimension of $\mathcal{T}_V(0, l)$ in Theorem 2.4.1: to find the dimension of $\mathcal{T}_V(0, l)$, one finds a basis first, and to do so one needs to define the tensor product. However, the tensor product of two differential forms (as two tensors) is not totally antisymmetric, and hence is no longer a differential form. Nonetheless, one can totally antisymmetrize all its indices and make it a differential form. Thus, we have the following definition:

Definition 2 Suppose ω and μ are respectively an l -form and an m -form, then their **wedge product** is an $(l + m)$ -form defined by the following equation:

$$(\omega \wedge \mu)_{a_1 \dots a_l b_1 \dots b_m} := \frac{(l+m)!}{l!m!} \omega_{[a_1 \dots a_l} \mu_{b_1 \dots b_m]} . \quad (5.1.2)$$

In other words, the wedge product is a map $\wedge : \Lambda(l) \times \Lambda(m) \rightarrow \Lambda(l+m)$ which satisfies (5.1.2).

The wedge product $(\omega \wedge \mu)_{a_1 \dots a_l b_1 \dots b_m}$ can also be denoted by $\omega_{a_1 \dots a_l} \wedge \mu_{b_1 \dots b_m}$, or $\omega \wedge \mu$ for short.

It follows from the definition that the wedge product satisfies both the associative law and distributive law, i.e., $(\omega \wedge \mu) \wedge v = \omega \wedge (\mu \wedge v)$ (and thus $\omega \wedge \mu \wedge v$ has a clear meaning) and $\omega \wedge (\mu + v) = \omega \wedge \mu + \omega \wedge v$. However, the wedge product does not in general obey the commutative law. For instance, for 1-forms ω and μ we have

$$\begin{aligned} \omega \wedge \mu &\equiv \omega_a \wedge \mu_b \equiv (\omega \wedge \mu)_{ab} = 2\omega_{[a}\mu_{b]} = \omega_a \mu_b - \omega_b \mu_a , \\ \mu \wedge \omega &\equiv (\mu \wedge \omega)_{ab} = 2\mu_{[a}\omega_{b]} = \mu_a \omega_b - \mu_b \omega_a , \end{aligned}$$

and thus for the wedge product of any two 1-forms we have $\omega \wedge \mu = -\mu \wedge \omega$. Carrying over to the general case, suppose ω and μ are an l - and an m -form, respectively, then

$$\omega \wedge \mu = (-1)^{lm} \mu \wedge \omega . \quad (5.1.3)$$

Theorem 5.1.3 Suppose $\dim V = n$, then

$$\dim \Lambda(l) = \frac{n!}{l!(n-l)!} , \quad \text{if } l \leq n ; \quad (5.1.4)$$

$$\Lambda(l) = \{0\} \text{ (only contains the zero element)} , \quad \text{if } l > n .$$

Proof Take $n = 3, l = 2$ as an example. Suppose $\{(e_1)^a, (e_2)^a, (e_3)^a\}$ is a basis of V , and $\{(e^1)_a, (e^2)_a, (e^3)_a\}$ is the corresponding dual basis, then ω_{ab} (as a tensor on V) can be expanded as

$$\begin{aligned}\omega_{ab} &= \omega_{11}(e^1)_a(e^1)_b + \omega_{12}(e^1)_a(e^2)_b + \omega_{13}(e^1)_a(e^3)_b \\ &\quad + \omega_{21}(e^2)_a(e^1)_b + \omega_{22}(e^2)_a(e^2)_b + \omega_{23}(e^2)_a(e^3)_b \\ &\quad + \omega_{31}(e^3)_a(e^1)_b + \omega_{32}(e^3)_a(e^2)_b + \omega_{33}(e^3)_a(e^3)_b.\end{aligned}$$

Noticing that $\omega_{11} = \omega_{22} = \omega_{33} = 0$, $\omega_{21} = -\omega_{12}$, $\omega_{32} = -\omega_{23}$, $\omega_{13} = -\omega_{31}$, the above equation becomes

$$\begin{aligned}\omega_{ab} &= \omega_{12}[(e^1)_a(e^2)_b - (e^2)_a(e^1)_b] + \omega_{23}[(e^2)_a(e^3)_b - (e^3)_a(e^2)_b] \\ &\quad + \omega_{31}[(e^3)_a(e^1)_b - (e^1)_a(e^3)_b] \\ &= \omega_{12}(e^1)_a \wedge (e^2)_b + \omega_{23}(e^2)_a \wedge (e^3)_b + \omega_{31}(e^3)_a \wedge (e^1)_b.\end{aligned}\quad (5.1.5)$$

Thus, any $\omega_{ab} \in \Lambda(2)$ can be expressed linearly in terms of $\{(e^1)_a \wedge (e^2)_b, (e^2)_a \wedge (e^3)_b, (e^3)_a \wedge (e^1)_b\}$. It is not difficult to show that the three 2-forms in the curly brackets are linearly independent (Exercise 5.1), and hence they comprise a set of basis vectors. Therefore, $\dim \Lambda(2) = 3$. The reader may generalize the above discussion to the case where l, n are arbitrary positive integers and $l \leq n$, and show that any l -form ω can be expanded as

$$\omega_{a_1 \dots a_l} = \sum_C \omega_{\mu_1 \dots \mu_l} (e^{\mu_1})_{a_1} \wedge \dots \wedge (e^{\mu_l})_{a_l}, \quad (5.1.6)$$

where $\{(e^1)_a, \dots, (e^n)_a\}$ is an arbitrary basis of V^* , and $\omega_{\mu_1 \dots \mu_l}$ are the components of ω in the basis of $\mathcal{T}_V(0, l)$ constituted by $\{(e^1)_a, \dots, (e^n)_a\}$, i.e.,

$$\omega_{\mu_1 \dots \mu_l} = \omega_{a_1 \dots a_l} (e_{\mu_1})^{a_1} \cdots (e_{\mu_l})^{a_l}, \quad (5.1.7)$$

\sum_C stands for summing over all combinations of taking l numbers from n numbers $(1, \dots, n)$, i.e., there are in total C_n^l vectors in the basis of $\Lambda(l)$, and hence we obtain (5.1.4). As for the case of $l > n$, it can be easily seen from Theorem 5.1.2 (b) that all the components of $\omega \in \Lambda(l)$ in this case are 0, and then $\Lambda(l)$ has only one element, namely the zero element: $\Lambda(l) = \{0\}$. \square

Equation (5.1.5) is a special case of (5.1.6) when $n = 3$ and $l = 2$. To make it easier to understand, here we provide another example: suppose $n = 4$ and $l = 3$, then (5.1.6) appears as

$$\begin{aligned}\omega_{abc} &= \omega_{123}(e^1)_a \wedge (e^2)_b \wedge (e^3)_c + \omega_{124}(e^1)_a \wedge (e^2)_b \wedge (e^4)_c \\ &\quad + \omega_{134}(e^1)_a \wedge (e^3)_b \wedge (e^4)_c + \omega_{234}(e^2)_a \wedge (e^3)_b \wedge (e^4)_c,\end{aligned}$$

where each component is determined by (5.1.7), e.g., $\omega_{134} = \omega_{abc}(e_1)^a(e_3)^b(e_4)^c$.

[Optional Reading 5.1.1]

Equation (5.1.6) can also be expressed as

$$\omega_{a_1 \dots a_l} = \frac{1}{l!} \omega_{\mu_1 \dots \mu_l} (e^{\mu_1})_{a_1} \wedge \dots \wedge (e^{\mu_l})_{a_l} \quad (\text{the symbol } \sum_{\mu_1, \dots, \mu_l}^n \text{ is omitted by convention}). \quad (5.1.6')$$

The number of nonzero terms on the right-hand side is equal to the number of permutations of taking l numbers from n numbers, i.e., $P_n^l = n!/(n-l)!$, which can be divided into $C_n^l = n!/[l!(n-l)!]$ groups, each containing $l!$ terms. All the terms in each group are the same, so dividing by $l!$ yields $C_n^l = n!/[l!(n-l)!]$ terms, which is in agreement with (5.1.6).

[The End of Optional Reading 5.1.1]

Now let us get back to a manifold M . If we assign an l -form on V_p to each point p on M (or $A \subset M$), we obtain an l -form field (the word “field” is usually omitted) on M (or A). 1-form fields and 0-form fields are simply dual vector fields and scalar fields, respectively. A smooth l -form field on M is called a **differential l -form**, also called an **l -form field** or an **l -form** for short.

Suppose (O, ψ) is a coordinate system, then an l -form field on O can be conveniently expressed pointwise linearly using a dual coordinate basis field $\{(dx^\mu)_a\}$. Setting $(e^\mu)_a$ in (5.1.6) to be $(dx^\mu)_a$, we have

$$\omega_{a_1 \dots a_l} = \sum_C \omega_{\mu_1 \dots \mu_l} (dx^{\mu_1})_{a_1} \wedge \dots \wedge (dx^{\mu_l})_{a_l}, \quad (5.1.8)$$

where

$$\omega_{\mu_1 \dots \mu_l} = \omega_{a_1 \dots a_l} (\partial/\partial x^{\mu_1})^{a_1} \dots (\partial/\partial x^{\mu_l})^{a_l} \quad (5.1.9)$$

is a function on O . An important special case is when $l = n$. Since now $C_n^l = C_n^n = 1$, there is only one term in the summation of (5.1.8), i.e.,

$$\omega_{a_1 \dots a_l} = \omega_{1 \dots n} (dx^1)_{a_1} \wedge \dots \wedge (dx^n)_{a_n}, \quad (5.1.10)$$

which can be shortened as

$$\omega = \omega_{1 \dots n} dx^1 \wedge \dots \wedge dx^n. \quad (5.1.10')$$

The equation above can be interpreted like this: the collection of all the n -forms at any point p in M is a 1-dimensional vector space, which only has one independent basis vector. Take the basis vector to be $dx^1 \wedge \dots \wedge dx^n|_p$, then (5.1.10') is the expansion of $\omega|_p$ in this basis. Note that the coefficient $\omega_{1 \dots n}$ can be different from point to point, and thus is a function on the coordinate patch, which can be expressed as a function of n variables, namely $\omega_{1 \dots n}(x^1, \dots, x^n)$.

We will use $\Lambda_M(l)$ to represent the collection of all the l -forms on M .

Definition 3 The exterior differentiation operator on a manifold M is the map $d : \Lambda_M(l) \rightarrow \Lambda_M(l+1)$, which can be defined as

$$(d\omega)_{ba_1 \dots a_l} := (l+1) \nabla_{[b} \omega_{a_1 \dots a_l]}, \quad (5.1.11)$$

where ∇_b is an arbitrary torsion-free derivative operator¹ (since it can be shown from $C^c_{ab} = C^c_{ba}$ that for arbitrary ∇ and $\tilde{\nabla}$ we have $\tilde{\nabla}_{[b}\omega_{a]} = \nabla_{[b}\omega_{a]}$). Fundamentally, one does not have to assign a derivative operator (or any additional structure, e.g., a metric) to M before defining the exterior differentiation operator.

Example 1 We have defined $(df)_a$ in Sect. 2.3, and we also know from (3.1.1) that $(df)_a = \nabla_a f$. Thus, $(df)_a$ is the exterior differentiation of $f \in \Lambda_M(0)$. This is exactly the reason why we used the symbol df .

An advantage of writing an l -form field ω in terms of the dual coordinate basis expansion (5.1.8) is the convenience of computing $d\omega$. See the following theorem:

Theorem 5.1.4 Suppose $\omega_{a_1 \dots a_l} = \sum_C \omega_{\mu_1 \dots \mu_l} (dx^{\mu_1})_{a_1} \wedge \dots \wedge (dx^{\mu_l})_{a_l}$, then

$$(d\omega)_{ba_1 \dots a_l} = \sum_C (\partial\omega_{\mu_1 \dots \mu_l})_b \wedge (dx^{\mu_1})_{a_1} \wedge \dots \wedge (dx^{\mu_l})_{a_l}. \quad (5.1.12)$$

Proof Exercise 5.4. Hint: choose the ordinary derivative operator ∂_a of this coordinate system as the ∇_b in (5.1.11). \square

Theorem 5.1.5 $d \circ d = 0$.

Proof Choosing the ordinary derivative operator ∂_a of an arbitrary coordinate system as the ∇_b in (5.1.11) yields

$$[d(d\omega)]_{cba_1 \dots a_l} = (l+2)(l+1)\partial_{[c}\partial_{[b}\omega_{a_1 \dots a_l]} = (l+2)(l+1)\partial_{[[c}\partial_{b]}\omega_{a_1 \dots a_l]} = 0,$$

where Theorem 2.6.2 (b) is used in the second equality, and $\partial_{[a}\partial_{b]}T^{\dots} = 0$ in Sect. 3.1 is used in the third equality. \square

Definition 4 Suppose ω is an l -form field on M . ω is said to be **closed** if $d\omega = 0$; ω is said to be **exact** if there exists an $(l-1)$ -form field μ such that $\omega = d\mu$.

Remark 1 Theorem 5.1.5 can be expressed alternatively as follows: if ω is exact, then ω is closed. However, to make the converse to be true one has to impose an additional requirement on M . The requirement is omitted here; what the reader has to know is that the trivial manifold \mathbb{R}^n satisfies this requirement. Since any manifold is locally trivial, one concludes that a closed l -form field on any manifold must be at least locally exact. That is, suppose ω is a closed l -form field on a manifold M , then for any point p of M there must be a neighborhood N on which there exists an $(l-1)$ -form field μ such that $\omega = d\mu$.

Corollary 5.1.6 When $M = \mathbb{R}^2$, Theorem 5.1.5 and its converse gives the following proposition in standard calculus: given functions $X(x, y)$ and $Y(x, y)$, a necessary and sufficient condition for the existence of a function $f(x, y)$ such that $df = Xdx + Ydy$ is $\partial X/\partial y = \partial Y/\partial x$.

¹ This definition is sufficient for this text, but the general definition of the exterior differentiation does not require the torsion-free condition, see, e.g., Warner (1983); Chern et al. (1999).

Proof It follows from Theorem 5.1.4 that the exterior differentiation of the 1-form field $Xdx + Ydy$ is

$$\begin{aligned} d(Xdx + Ydy) &= dX \wedge dx + dY \wedge dy \\ &= \left(\frac{\partial X}{\partial x} dx + \frac{\partial X}{\partial y} dy \right) \wedge dx + \left(\frac{\partial Y}{\partial x} dx + \frac{\partial Y}{\partial y} dy \right) \wedge dy \\ &= \frac{\partial X}{\partial y} dy \wedge dx + \frac{\partial Y}{\partial x} dx \wedge dy = \left(\frac{\partial Y}{\partial x} - \frac{\partial X}{\partial y} \right) dx \wedge dy. \end{aligned} \quad (5.1.13)$$

(A) If there exists a function f such that the equality $df = Xdx + Ydy$ of 1-form fields holds, then it follows from (5.1.13) that

$$\left(\frac{\partial Y}{\partial x} - \frac{\partial X}{\partial y} \right) dx \wedge dy = dd^f = 0.$$

Hence, $\partial X/\partial y = \partial Y/\partial x$.

(B) If $\partial X/\partial y = \partial Y/\partial x$, then it follows from (5.1.13) that $d(Xdx + Ydy) = 0$, namely the 1-form field $Xdx + Ydy$ is closed. Therefore, $Xdx + Ydy$ is exact (because $M = \mathbb{R}^2$), i.e., there exists a function f such that $Xdx + Ydy = df$. \square

[Optional Reading 5.1.2]

When we say a property holds locally on a manifold M , we mean $\forall p \in M \exists$ a neighborhood N of p such that this property holds on N . What is important is that $\forall p \in M$ there is such an N ; thus, “holds locally” does not mean it only holds in a local area but not anywhere else. The crucial point of the word “local” is to emphasize it does not necessarily hold (globally) on the whole manifold M . We hereby give three examples to help the reader understand this.

1. People often hear that “any manifold looks locally like \mathbb{R}^n ”, the precise meaning is that: *every point* p of M has a coordinate neighborhood O such that there exists a homeomorphism (which can be promoted to a diffeomorphism) $\psi : O \rightarrow \psi[O] \subset \mathbb{R}^n$, and thus, O and $\psi[O]$ “cannot be more alike”. One can always choose an O such that $\psi[O]$ is homeomorphic to \mathbb{R}^n , and hence M looks locally like \mathbb{R}^n . However, M may not look globally like \mathbb{R}^n , i.e., there may not exist a diffeomorphism from M to \mathbb{R}^n .

2. “A closed l -form field is locally exact” means that $\forall p \in M \exists$ a neighborhood N of p , there is an $(l-1)$ -form field μ such that $\omega = d\mu$ on N . However, there may not exist a global $(l-1)$ -form field μ on M that satisfies $\omega = d\mu$.

3. “A Möbius strip (see Fig. 5.3) looks locally like C^2 (a cylinder)” means that $\forall p \in M \exists$ an open neighborhood N of p such that N is diffeomorphic to an open subset of C^2 . However, there does not exist a diffeomorphism from the whole Möbius strip to C^2 .

The properties involved in the three examples above all hold locally, which demonstrates the importance of distinguishing local properties from global properties.

[The End of Optional Reading 5.1.2]

5.2 Integration on Manifolds

First, we take the 3-dimensional Euclidean space $(\mathbb{R}^3, \delta_{ab})$ as an example. Suppose \vec{v} is a vector field, L is a smooth curve, and S is a smooth surface. Before we specify

Fig. 5.1 A curve in Euclidean space. The arrow represents the assigned direction of the integral



Fig. 5.2 A surface in Euclidean space. \vec{n} is the assigned normal direction

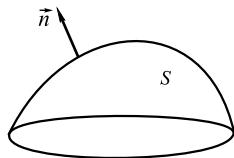


Fig. 5.3 Möbius strip (an example of a non-orientable manifold)



the direction of L (the arrow in Fig. 5.1) and the normal direction of S (the arrow \vec{n} in Fig. 5.2), both the integrals $\int_L \vec{v} \cdot d\vec{l}$ and $\iint_S \vec{v} \cdot d\vec{S}$ can only be determined up to a minus sign. By extension, one should assign an “orientation” to a manifold before calculating the integral on it. However, not all manifolds are orientable.

Definition 1 An n -dimensional manifold is said to be **orientable** if there exists on it a C^0 nowhere vanishing n -form field ε .

Example 1 \mathbb{R}^3 is an orientable manifold, since there exists a C^∞ 3-form field $\varepsilon \equiv dx \wedge dy \wedge dz$ on \mathbb{R}^3 , where x, y, z are natural coordinates.

Example 2 A Möbius strip is a non-orientable manifold (Fig. 5.3).

Definition 2 If a C^0 nowhere vanishing n -form field ε is given on an n -dimensional orientable manifold M , then we say M is **oriented**. Suppose ε_1 and ε_2 are two different C^0 nowhere vanishing n -form fields. If there exists a function h that is positive everywhere such that $\varepsilon_1 = h\varepsilon_2$, then we say ε_1 and ε_2 provide the same **orientation** to M .

Remark 1 From the orientation point of view, the ε_1 and ε_2 that satisfy $\varepsilon_1 = h\varepsilon_2$ ($h > 0$) are equivalent. Since the collection of all the 1-forms at each point on an n -dimensional manifold M is a 1-dimensional vector space (see (5.1.4)), for any two n -form fields ε_1 and ε_2 there must be $\varepsilon_1 = h\varepsilon_2$, where h is a (not necessarily positive) function on M . If ε_1 and ε_2 are nowhere vanishing, then h is nowhere

vanishing; if ϵ_1 and $h\epsilon_2$ are C^0 , then h is C^0 . For a connected manifold² (we will only talk about connected manifolds), a nowhere vanishing function can only be either positive everywhere or negative everywhere. Thus, a connected manifold can only have two kinds of orientations.

Definition 3 After we choose an orientation on M represented by ϵ , a basis field $\{(e_\mu)^a\}$ on an open subset $O \subset M$ is said to be **right-handed** measured by ϵ if there exists a function h on O that is positive everywhere such that $\epsilon = h(e^1)_{a_1} \wedge \cdots \wedge (e^n)_{a_n}$, where $\{(e^\mu)_a\}$ is the dual basis of $\{(e_\mu)^a\}$ (otherwise it is said to be **left-handed**). A coordinate system is called a **right (left)-handed system** if its coordinate basis is right (left)-handed.

Now we introduce the integral of an n -form field ω on an n -dimensional oriented manifold M . ω can be expanded using the wedge product $dx^1 \wedge \cdots \wedge dx^n$ of a dual coordinate basis as [see (5.1.10')]

$$\omega = \omega_{1 \dots n}(x^1, \dots, x^n) dx^1 \wedge \cdots \wedge dx^n. \quad (5.2.1)$$

Thus, each n -form field ω gives rise to a function of n variables, i.e., $\omega_{1 \dots n}(x^1, \dots, x^n)$, in the coordinate patch. We call the n -tuple integral of this function of n variables the integral of the n -form field ω ; the precise definition is as follows:

Definition 4 Suppose (O, ψ) is a right-handed coordinate system on an n -dimensional oriented manifold M , ω is a continuous n -form field on an open subset $G \subset O$, then the **integral** of ω on G is defined as

$$\int_G \omega := \int_{\psi[G]} \omega_{1 \dots n}(x^1, \dots, x^n) dx^1 \cdots dx^n. \quad (5.2.2)$$

The right-hand side of the above equation is just the standard integral³ of a function of n variables on an open subset $\psi[G]$ of \mathbb{R}^n , which is already well-defined.

Remark 2 (1) To show the validity of Definition 4, one should also prove that the integral of ω on G does not depend on the choice of the right-handed system. We only prove the case $n = 2$ below as an example; the reader should carry over the proof to the general case.

Suppose (O, ψ) and (O', ψ') are right-handed coordinate systems that satisfy $G \subset O \cap O'$. The coordinates of these two systems are denoted by x^1, x^2 and x'^1 ,

² A topological space (X, \mathcal{T}) is said to be **connected** if it only has two subsets that are both open and closed (Definition 7 of Sect. 1.2), and is said to be **arcwise connected** if any two points in X can be joined by a continuous curve in X . A manifold is said to be **connected** (or **arcwise connected**) if its base topological space is **connected** (or **arcwise connected**). For a topological space, arcwise connected must be connected, but connected is not necessary arcwise connected (there exist “sideswipe” counterexamples). For a manifold, arcwise connected is equivalent to connected [see Abraham and Marsden (1978) Proposition 1.1.33].

³ Namely, the Riemann or Lebesgue integral.

x'^2 , respectively, then

$$\omega = \omega_{12} dx^1 \wedge dx^2 = \omega'_{12} dx'^1 \wedge dx'^2.$$

Let $\int_G \omega \equiv \int_{\psi[G]} \omega_{12} dx^1 dx^2$ and $(\int_G \omega)' \equiv \int_{\psi'[G]} \omega'_{12} dx'^1 dx'^2$. We want to prove

$$\left(\int_G \omega \right)' = \int_G \omega. \quad (5.2.3)$$

From the tensor transformation law we see that $\omega'_{12} = \frac{\partial x^1}{\partial x'^1} \frac{\partial x^2}{\partial x'^2} \omega_{12} + \frac{\partial x^2}{\partial x'^1} \frac{\partial x^1}{\partial x'^2} \omega_{21} = \omega_{12} \det\left(\frac{\partial x^\mu}{\partial x'^\nu}\right)$, where

$$\det\left(\frac{\partial x^\mu}{\partial x'^\nu}\right) \equiv \begin{vmatrix} \frac{\partial x^1}{\partial x'^1} & \frac{\partial x^1}{\partial x'^2} \\ \frac{\partial x^2}{\partial x'^1} & \frac{\partial x^2}{\partial x'^2} \end{vmatrix}$$

is the Jacobian of this coordinate transformation. According to a well-known law in multivariable calculus,

$$\int_{\psi[G]} \omega_{12} dx^1 dx^2 = \int_{\psi'[G]} \omega_{12} \det(\partial x^\mu / \partial x'^\nu) dx'^1 dx'^2 = \int_{\psi'[G]} \omega'_{12} dx'^1 dx'^2, \quad (5.2.4)$$

and hence (5.2.3) is proved.

However, if $\{x^\mu\}$ and $\{x'^\mu\}$ are right and left-handed systems, respectively, then we have $\det(\partial x^\mu / \partial x'^\nu) < 0$. From the multivariable calculus we know that the $\det(\partial x^\mu / \partial x'^\nu)$ on the right-hand side of the first equality in (5.2.4) should be changed to $|\det(\partial x^\mu / \partial x'^\nu)| = -\det(\partial x^\mu / \partial x'^\nu)$, and hence (5.2.4) turns to

$$\int_{\psi[G]} \omega_{12} dx^1 dx^2 = \int_{\psi'[G]} \omega_{12} \det(\partial x^\mu / \partial x'^\nu) dx'^1 dx'^2 = - \int_{\psi'[G]} \omega'_{12} dx'^1 dx'^2. \quad (5.2.5)$$

Therefore, to make sure the definition of the integral is consistent, when $\{x^\mu\}$ is left-handed $\int_G \omega$ should be defined as

$$\int_G \omega := - \int_{\psi[G]} \omega_{1\dots n}(x^1, \dots, x^n) dx^1 \cdots dx^n. \quad (5.2.6)$$

(2) Whether a coordinate system is right-handed or left-handed is determined by the orientation one chooses for the manifold. Hence, $\int_G \omega$ defined by (5.2.2) and (5.2.6) depends on the orientation given by ϵ , and the sign of the integral will change when the orientation is changed.

(3) Definition 4 only defines the integral of ω on an open subset G in the coordinate patch. The integral of ω over the whole manifold M can be defined by “sewing” the local integrals, which entails the concept of a “partition of unity”. [The reader may refer to Wald (1984).]

Suppose S and M are manifolds with dimensions l and $n(> l)$, respectively, and $\phi : S \rightarrow M$ is an embedding (see Sect. 4.4). Since $\phi[S]$ is an l -dimensional submanifold, of course we can talk about the integral of an l -form field μ on it (Definition 4 applies). However, the fact that “ $\phi[S]$ is embedded in M ” leads to two possible meanings of “an l -form field on $\phi[S]$ ”. Just like “a vector field on $\phi[S]$ ” can be tangent or not tangent to $\phi[S]$, “an l -form field on $\phi[S]$ ” can also be classified as “tangent to” and not “tangent to” $\phi[S]$. Precisely speaking, an l -form field μ on $\phi[S]$ is said to be “tangent to” $\phi[S]$ if $\forall q \in \phi[S], \mu|_q$ is an l -form on W_q (rather than V_q); that is, $\mu|_q$ is a linear map that can turn l arbitrary elements of W_q into a real number. An “ l -form field on $\phi[S]$ ” can either be tangent to $\phi[S]$ or not “tangent to” $\phi[S]$. Since we consider the $\phi[S]$ as an independent manifold when we talk about the integral of an l -form on $\phi[S]$ (and do not care about the “outside” situation), only an l -form μ that is “tangent to” $\phi[S]$ is meaningful. Nevertheless, since an l -form field μ on $\phi[S]$ that is not “tangent to” $\phi[S]$ is a linear map that can turn l arbitrary elements in V_q (rather than only W_q) of each point $q \in \phi[S]$ into a real number, and W_q is nothing but a subspace of V_q , we can obtain an l -form $\tilde{\mu}$ that is “tangent to” $\phi[S]$ by just restricting the acting range of μ to W_q . We denote it by $\tilde{\mu}$ and call it the **restriction of μ** . Precisely, we have the following definition:

Definition 5 Suppose $\mu_{a_1 \dots a_l}$ is an l -form field on an l -dimensional submanifold $\phi[S] \subset M$. An l -form field $\tilde{\mu}_{a_1 \dots a_l}$ on $\phi[S]$ (viewed as a manifold independent of M) is called the **restriction** of the l -form field $\mu_{a_1 \dots a_l}$ on $\phi[S]$ if

$$\begin{aligned} \tilde{\mu}_{a_1 \dots a_l}|_q(w_1)^{a_1} \cdots (w_l)^{a_l} &= \mu_{a_1 \dots a_l}|_q(w_1)^{a_1} \cdots (w_l)^{a_l}, \\ \forall q \in \phi[S], \quad (w_1)^{a_1} \cdots (w_l)^{a_l} &\in W_q. \end{aligned} \quad (5.2.7)$$

Similar to the induced metric (see Definition 5' of Sect. 4.4), in the perspective that a submanifold is the embedding map $\phi : S \rightarrow M$ itself, the restriction of a form μ is essentially the pullback $\phi^* \mu$ on S . Especially, one can show that the integral of the $\tilde{\mu}$ in Definition 5 satisfies

$$\int_{\phi[S]} \tilde{\mu} = \int_S \phi^* \mu.$$

Later on, whenever we talk about the integral of an l -form field μ over an l -dimensional submanifold $\phi[S]$, one should always interpret it as the integral of the restriction of μ , i.e., always interpret $\int_{\phi[S]} \mu$ as $\int_{\phi[S]} \tilde{\mu}$ or $\int_S \phi^* \mu$.

5.3 Stokes's Theorem

In the 3-dimensional Euclidean space, the Stokes theorem

$$\iint_S (\vec{\nabla} \times \vec{A}) \cdot d\vec{S} = \oint_L \vec{A} \cdot d\vec{l}$$

and Gauss's theorem

$$\iiint_V (\vec{\nabla} \cdot \vec{A}) dV = \iint_S \vec{A} \cdot \vec{n} dS$$

share a common property in that they manifest a relationship between an integral over a region and an integral on the boundary. Before we bring up the general Stokes theorem, we first introduce the concept of “a manifold with boundary”. The simplest example for an n -dimensional manifold with boundary is

$$\mathbb{R}^{n-} := \{(x^1, \dots, x^n) \in \mathbb{R}^n | x^1 \leq 0\},$$

where x^1, \dots, x^n are natural coordinates, the subset formed by all the points on $x^1 = 0$ is called the boundary of \mathbb{R}^{n-} , which by itself is an $(n-1)$ -dimensional manifold (in fact it is just \mathbb{R}^{n-1}). Carrying over to the general case, an **n -dimensional manifold N with boundary** is defined in a way similar to an n -dimensional manifold, except the \mathbb{R}^n in that definition is changed to \mathbb{R}^{n-} . That is, each element in the open cover $\{O_\alpha\}$ of N should be homeomorphic to an open subset of \mathbb{R}^{n-} ; all the points in N that are mapped to $x^1 = 0$ (such as p in Fig. 5.4) form the **boundary** of N , denoted by ∂N . Note that ∂N is an $(n-1)$ -dimensional manifold; $i(N) \equiv N - \partial N$ is an n -dimensional manifold. For instance, a solid ball B in \mathbb{R}^3 is a 3-dimensional manifold with boundary, whose boundary (a 2-sphere) is a 2-dimensional manifold, while $i(B)$ is a 3-dimensional manifold.

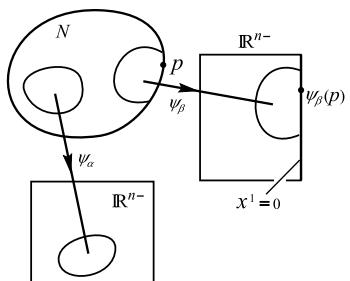
Theorem 5.3.1 (Stokes's Theorem) *Suppose a compact subset N of an n -dimensional oriented manifold is an n -dimensional manifold with boundary, and ω is an $(n-1)$ -form field (whose differentiability is at least C^1) on M , then*

$$\int_{i(N)} d\omega = \int_{\partial N} \omega. \quad (5.3.1)$$

Proof See, for example, Chern et al. (1999). □

Remark 1 Restricting the orientation ϵ of M on N yields the orientation of N , also denoted by ϵ , which naturally induces an orientation on the boundary ∂N of N , denoted by $\bar{\epsilon}$, short for $\bar{\epsilon}_{a_1 \dots a_{n-1}}$. Take \mathbb{R}^{2-} as an example, in which $M = \mathbb{R}^2$,

Fig. 5.4 A diagrammatic sketch for a manifold N with boundary, in which p is a boundary point



$N = \mathbb{R}^{2-}$, $\partial N = \{(x^1, x^2) | x^1 = 0\}$. Suppose the orientation of \mathbb{R}^2 (and, consequently, \mathbb{R}^{2-}) is $\varepsilon_{ab} = (dx^1)_a \wedge (dx^2)_b$, then $\{x^1, x^2\}$ is a right-handed system measured by ε_{ab} . Since $x^1|_{\partial N} = 0$, after getting rid of x^1 , $\{x^2\}$ is a coordinate system of ∂N . We define $\bar{\varepsilon}_a$ as the induced orientation of ∂N such that $\{x^2\}$ is a right-handed system measured by $\bar{\varepsilon}_a$. This requirement can be satisfied by choosing $\bar{\varepsilon}_a = (dx^2)_a$. This basic requirement of an induced orientation can be generalized to any manifold N with boundary [for details, see Wald (1984) p. 431]. The left-hand side of (5.3.1) is the integral of an n -form field $d\omega$ over an n -dimensional manifold $i(N)$ (with ε as its orientation), and the right-hand side is the integral of an $(n-1)$ -form field ω over an $(n-1)$ -dimensional manifold ∂N (with $\bar{\varepsilon}$ as its orientation).

Example 1 Suppose \vec{A} is a vector field on the 2-dimensional Euclidean space, L is a smooth closed curve in \mathbb{R}^2 , S is an open subset surrounded by L (see Fig. 5.5), x^1 and x^2 are Cartesian coordinates. Then, the familiar Stokes theorem for 2-dimensional Euclidean space (also called Green's theorem) is

$$\iint_S (\partial A_2 / \partial x^1 - \partial A_1 / \partial x^2) dx^1 dx^2 = \oint_L A_l dl. \quad (5.3.2)$$

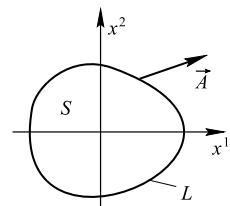
Now we will show that the equation above is a special case of Theorem 5.3.1. Let $M = \mathbb{R}^2$, then $S \cup L$ can be treated as the N in Theorem 5.3.1, where S and L serve as $i(N)$ and ∂N , respectively. If we turn A^a into a 1-form field using the Euclidean metric δ_{ab} , then A_a can be treated as the ω in Theorem 5.3.1. Expand A_a using the dual coordinate basis vectors of the Cartesian system: $\omega = A_a = A_\mu (dx^\mu)_a$, then

$$\begin{aligned} d\omega = dA_\mu \wedge dx^\mu &= \frac{\partial A_\mu}{\partial x^\nu} dx^\nu \wedge dx^\mu = \frac{\partial A_1}{\partial x^2} dx^2 \wedge dx^1 + \frac{\partial A_2}{\partial x^1} dx^1 \wedge dx^2 \\ &= \left(\frac{\partial A_2}{\partial x^1} - \frac{\partial A_1}{\partial x^2} \right) dx^1 \wedge dx^2. \end{aligned}$$

Thus, the left-hand side of (5.3.2) can be expressed as $\int_{i(N)} d\omega$, which means it is a special case of the left-hand side of (5.3.1). On the other hand, the right-hand side of (5.3.1) is $\int_{\partial N} \omega = \int_{\partial N} \tilde{\omega}$. Setting the arc length l as the local coordinate of L , expanding $\tilde{\omega}$ using the dual coordinate basis vector as $\tilde{\omega}_a = \tilde{\omega}_1(l) (dl)_a$, and contracting both sides with $(\partial/\partial l)^a$, we have

$$\tilde{\omega}_1(l) = \tilde{\omega}_a (\partial/\partial l)^a = \omega_a (\partial/\partial l)^a = A_a (\partial/\partial l)^a = A_l,$$

Fig. 5.5 Stokes's theorem for 2-dimensional Euclidean space



and hence $\tilde{\omega} = A_l dl$. Therefore, the right-hand side of (5.3.2) can be written as

$$\oint_L A_l dl = \int_{\partial N} \omega. \quad (5.3.3)$$

Thus, we can see that (5.3.2) is a special case of (5.3.1).

[Optional Reading 5.3.1]

There is one thing we need to make clear in the derivation of (5.3.3). The integral region of $\int_L \tilde{\omega}$ is a closed curve L , which is a 1-dimensional non-trivial manifold that takes at least two coordinate patches to cover it. Therefore, one should perform the local integral for each coordinate patch and then “sew” them together. Luckily, now we can deal with it in a simple way: suppose L' is the manifold coming from removing a point from L , then L' can be covered by one coordinate patch, and since it does not affect the value of the integral when a point is removed, the derivation is valid.

[The End of Optional Reading 5.3.1]

Now we have introduced the integral of a differential form on a manifold and some related theorems. To talk about the integral of a function on a manifold, we first introduce the concept of a volume element in Sect. 5.4.

5.4 Volume Elements

Definition 1 An arbitrary C^0 and nowhere vanishing n -form field ϵ on an n -dimensional orientable manifold M is called a **volume element**.

Remark 1 The difference between a volume element and an orientation is that: if ϵ_1 and ϵ_2 are two C^0 and nowhere vanishing n -form fields, and there is a function h that is positive everywhere such that $\epsilon_1 = h\epsilon_2$, then ϵ_1 and ϵ_2 represent the same orientation; however, as long as $\epsilon_1 \neq \epsilon_2$, they are two different volume elements. For an orientable connected manifold, there are only two orientations, while there are infinitely many volume elements. When talking about integration or the volume elements on an orientable manifold, one does not need to assign a metric field to the manifold. The choice of a volume element is quite arbitrary, and no volume element is special. (There is only one requirement: the volume element has to be compatible with the orientation, i.e., the multiplicative factor between the ϵ representing the volume element and the ϵ representing the orientation is positive.) However, if a metric field g_{ab} is assigned to the manifold, then there exists a natural way to choose a specific volume element.

First we consider a 2-dimensional oriented manifold with a metric g_{ab} . Suppose $\epsilon_{a_1 a_2}$ is an arbitrary volume element, then $\epsilon^{a_1 a_2} \equiv g^{a_1 b_1} g^{a_2 b_2} \epsilon_{b_1 b_2}$ is meaningful, and $\epsilon^{a_1 a_2} \epsilon_{a_1 a_2}$ is a scalar field that can be computed using any basis. Choose the orthonormal basis. If g_{ab} is a positive definite metric, then

$$\epsilon^{a_1 a_2} \epsilon_{a_1 a_2} = \delta^{\mu_1 \nu_1} \delta^{\mu_2 \nu_2} \epsilon_{\nu_1 \nu_2} \epsilon_{\mu_1 \mu_2} = \delta^{11} \delta^{22} \epsilon_{12} \epsilon_{12} + \delta^{22} \delta^{11} \epsilon_{21} \epsilon_{21} = 2(\epsilon_{12})^2.$$

If g_{ab} is Lorentzian, then

$$\varepsilon^{a_1 a_2} \varepsilon_{a_1 a_2} = \eta^{11} \eta^{22} \varepsilon_{12} \varepsilon_{12} + \eta^{22} \eta^{11} \varepsilon_{21} \varepsilon_{21} = -2(\varepsilon_{12})^2.$$

Generalizing to an n -dimensional manifold with an arbitrary metric g_{ab} we have

$$\varepsilon^{a_1 \dots a_n} \varepsilon_{a_1 \dots a_n} = (-1)^s n! (\varepsilon_{1 \dots n})^2,$$

where $\varepsilon_{1 \dots n}$ is a component of $\varepsilon_{a_1 \dots a_n}$ in the orthonormal basis, and s is the number of -1 among the components of g_{ab} in the orthonormal basis; for instance, $s = 0$ for definite positive metrics, and $s = 1$ for Lorentzian metrics. To choose a specific volume element using the given metric, one just needs to impose the following simple requirement on the components of the volume element $\varepsilon_{a_1 \dots a_n}$ in the orthonormal basis $\{(e^\mu)_a\}$:

$$\varepsilon_{1 \dots n} = \pm 1, \quad (5.4.1)$$

i.e.,

$$\varepsilon_{a_1 \dots a_n} = \pm (e^1)_{a_1} \wedge \dots \wedge (e^n)_{a_n} \quad (\text{for an orthonormal basis}), \quad (5.4.2)$$

which is equivalent to requiring

$$\varepsilon^{a_1 \dots a_n} \varepsilon_{a_1 \dots a_n} = (-1)^s n!. \quad (5.4.3)$$

An $\varepsilon_{a_1 \dots a_n}$ that satisfies the above equation is called the **volume element associated (or compatible) with the metric g_{ab}** . The above equation can only determine the volume element up to a minus sign, only together with the requirement “the volume element is compatible with the orientation” can the volume element be uniquely determined. Thus, the $+$ and $-$ signs on the right-hand side of (5.4.2) correspond to right and left-handed orthonormal bases.

Summary. When dealing with an integral, we are only concerned here with orientable manifolds.⁴ First, one should choose an orientation and make M an orientable manifold. A basis being right or left-handed is stipulated by the orientation we choose. When there is no metric field g_{ab} (or any other available geometric structure), except for being required to be compatible with the orientation, the volume element is quite arbitrary. After g_{ab} is assigned, $\varepsilon_{a_1 \dots a_n}$ is uniquely determined by g_{ab} and the requirement of it being compatible with the orientation, called the **associated volume element** for short. Later on, unless stated otherwise, all the volume elements we mention when there is a metric will refer to this unique associated volume element.

Choose any right-handed Cartesian system $\{x, y, z\}$ in the 3-dimensional Euclidean space $(\mathbb{R}^3, \delta_{ab})$ by intuition and assign the orientation using the 3-form field $\boldsymbol{\varepsilon} = dx \wedge dy \wedge dz$, then according to Definition 3 of Sect. 5.2, $\{x, y, z\}$ is a right-handed system measured by $\boldsymbol{\varepsilon}$. Comparing $\boldsymbol{\varepsilon} = dx \wedge dy \wedge dz$ and (5.4.2) we can see that $\boldsymbol{\varepsilon}$ is an associated volume element. Suppose G is an open subset of \mathbb{R}^3

⁴ Integration can also be defined on non-orientable manifolds. In this case, one needs the concept of a “twisted” (also called “odd” or “pseudo”) form, which is outside the scope of this text.

and the integral $\iiint_G dx dy dz$ exists, then this integral naturally stands for the volume of G (by the definition of volume in standard calculus). On the other hand, it follows from Definition 4 of Sect. 5.2 that the integral $\int_G \epsilon$ of the 3-form field ϵ on $G \subset \mathbb{R}^3$ is exactly $\iiint_G dx dy dz$, and thus $\int_G \epsilon$ is the volume of G . Generalize to any oriented manifold N with a positive definite metric g_{ab} : suppose ϵ is the associated volume element, if $\int_N \epsilon$ exists, then we call it the **volume** (or **length** and **area** for 1- and 2-dimensional manifolds, respectively) of N (measured by g_{ab}). This is the reason why ϵ is called a volume element.

Theorem 5.4.1 Suppose ϵ is an associated volume element, $\{(e_\mu)^a\}$ and $\{(e^\mu)_a\}$ are a basis and its dual basis, g is the determinant of the components of g_{ab} in this basis, $|g|$ is the absolute value of g , then (+ for right-handed basis and - for left-handed basis)

$$\epsilon_{a_1 \dots a_n} = \pm \sqrt{|g|} (e^1)_{a_1} \wedge \dots \wedge (e^n)_{a_n}. \quad (5.4.4)$$

Proof [Optional Reading]

From (5.4.3) we know that ϵ and the components of g_{ab} in the given basis satisfy

$$(-1)^s n! = \epsilon^{\mu_1 \dots \mu_n} \epsilon_{\mu_1 \dots \mu_n} = g^{\mu_1 \nu_1} \dots g^{\mu_n \nu_n} \epsilon_{\nu_1 \dots \nu_n} \epsilon_{\mu_1 \dots \mu_n}. \quad (5.4.5)$$

The right-hand side of this equation should be interpreted as summing over each of $\mu_1 \dots \mu_n$ and $\nu_1 \dots \nu_n$ from 1 to n . Considering the total antisymmetry of $\epsilon_{\nu_1 \dots \nu_n}$ and $\epsilon_{\mu_1 \dots \mu_n}$, one can simplify the summation above into a sum over the permutations. More precisely, let $\sum_{\pi(\nu_1 \dots \nu_n)}$ represent summing over all the permutations of $1, 2, \dots, n$, then

$$\begin{aligned} \text{r.h.s. of (5.4.5)} &= \sum_{\pi(\mu_1 \dots \mu_n)} \sum_{\pi(\nu_1 \dots \nu_n)} g^{\mu_1 \nu_1} \dots g^{\mu_n \nu_n} \epsilon_{\nu_1 \dots \nu_n} \epsilon_{\mu_1 \dots \mu_n} \\ &= \sum_{\pi(\mu_1 \dots \mu_n)} g^{\mu_1 1} g^{\mu_2 2} g^{\mu_3 3} \dots g^{\mu_n n} \epsilon_{123 \dots n} \epsilon_{\mu_1 \dots \mu_n} \\ &\quad + \sum_{\pi(\mu_1 \dots \mu_n)} g^{\mu_1 2} g^{\mu_2 1} g^{\mu_3 3} \dots g^{\mu_n n} \epsilon_{213 \dots n} \epsilon_{\mu_1 \dots \mu_n} + \dots. \end{aligned} \quad (5.4.5')$$

There are $n!$ terms on the right-hand side of this equation. Using $\hat{\epsilon}_{\mu_1 \dots \mu_n}$ to represent the Levi-Civita symbol, i.e.,

$$\hat{\epsilon}_{\mu_1 \dots \mu_n} = \begin{cases} +1, & (\text{when } \mu_1 \dots \mu_n \text{ is an even permutation of } 1, 2, \dots, n), \\ -1, & (\text{when } \mu_1 \dots \mu_n \text{ is an odd permutation of } 1, 2, \dots, n), \\ 0, & (\text{when two of } \mu_1, \dots, \mu_n \text{ are equal}), \end{cases}$$

we have $\epsilon_{\mu_1 \dots \mu_n} = \epsilon_{123 \dots n} \hat{\epsilon}_{\mu_1 \dots \mu_n}$. Denote $\sum_{\pi(\mu_1 \dots \mu_n)}$ as \sum_{π} for short, then

the first term on the r.h.s. of (5.4.5')

$$= (\epsilon_{123 \dots n})^2 \sum_{\pi} g^{\mu_1 1} g^{\mu_2 2} g^{\mu_3 3} \dots g^{\mu_n n} \hat{\epsilon}_{\mu_1 \mu_2 \mu_3 \dots \mu_n} = (\epsilon_{123 \dots n})^2 \det(g^{\mu\nu}),$$

where $\det(g^{\mu\nu})$ stands for the determinant of the matrix $g^{\mu\nu}$ (the definition of the determinant is used in the last step). Also,

the second term on the r.h.s. of (5.4.5')

$$= - \sum_{\pi} g^{\mu_1 2} g^{\mu_2 1} g^{\mu_3 3} \dots g^{\mu_n n} \epsilon_{123 \dots n} \epsilon_{\mu_1 \mu_2 \mu_3 \dots \mu_n}$$

$$\begin{aligned}
&= -(\varepsilon_{123\dots n})^2 \sum_{\pi} g^{\mu_1 2} g^{\mu_2 1} g^{\mu_3 3} \dots g^{\mu_n n} \hat{\varepsilon}_{\mu_1 \mu_2 \mu_3 \dots \mu_n} \\
&= -(\varepsilon_{1\dots n})^2 \sum_{\pi} g^{\mu_2 2} g^{\mu_1 1} g^{\mu_3 3} \dots g^{\mu_n n} \hat{\varepsilon}_{\mu_2 \mu_1 \mu_3 \dots \mu_n} \\
&= (\varepsilon_{1\dots n})^2 \sum_{\pi} g^{\mu_1 1} g^{\mu_2 2} g^{\mu_3 3} \dots g^{\mu_n n} \hat{\varepsilon}_{\mu_1 \mu_2 \mu_3 \dots \mu_n} \\
&= (\varepsilon_{123\dots n})^2 \det(g^{\mu\nu}).
\end{aligned}$$

Similarly one can prove that each term on the right-hand side of (5.4.5') equals $(\varepsilon_{1\dots n})^2 \det(g^{\mu\nu})$. Noticing that there are $n!$ terms on the right-hand side of the above equation, plugging them back to (5.4.5) yields $(-1)^s n! = (n!) (\varepsilon_{1\dots n})^2 \det(g^{\mu\nu})$, or $(-1)^s = (\varepsilon_{1\dots n})^2 \det(g^{\mu\nu})$. The fact that the matrix $g^{\mu\nu}$ is the inverse of $g_{\mu\nu}$ gives that $\det(g^{\mu\nu}) = 1/\det(g_{\mu\nu}) \equiv 1/g$. Plugging into the previous equation, we obtain

$$(-1)^s g = (\varepsilon_{1\dots n})^2, \quad \varepsilon_{1\dots n} = \pm \sqrt{|g|},$$

and therefore we have (5.4.4). \square

Remark 2 For an orthonormal basis we have $|g| = 1$, and hence (5.4.4) goes back to (5.4.2).

Theorem 5.4.2 Suppose ∇_a and ε are respectively the derivative operator and the volume element associated with the metric, then

$$\nabla_b \varepsilon_{a_1 \dots a_n} = 0. \quad (5.4.6)$$

Proof It follows from $\nabla_b g_{ac} = 0$ and (5.4.3) that $\varepsilon^{a_1 \dots a_n} \nabla_b \varepsilon_{a_1 \dots a_n} = 0$, and thus for any vector field v^b we have

$$\varepsilon^{a_1 \dots a_n} v^b \nabla_b \varepsilon_{a_1 \dots a_n} = 0. \quad (5.4.7)$$

Since the collection of all the n -forms at a point in M is a 1-dimensional vector space, any two n -forms at this point can only differ by a multiplicative factor h (h can be different from point to point). Therefore, $v^b \nabla_b \varepsilon_{a_1 \dots a_n} = h \varepsilon_{a_1 \dots a_n}$. Plugging into (5.4.7) gives $h = 0$, and thus $v^b \nabla_b \varepsilon_{a_1 \dots a_n} = 0$. Since v^b is an arbitrary vector field, we have $\nabla_b \varepsilon_{a_1 \dots a_n} = 0$. \square

Now we are going to prove two identities about volume elements that are quite useful. To do so, we need to prove the following lemma first.

Lemma 5.4.3

$$\delta^{[a_1}_{a_1} \dots \delta^{a_j}_{a_j} \delta^{a_{j+1}}_{b_{j+1}} \dots \delta^{a_n]}_{b_n} = \frac{(n-j)! j!}{n!} \delta^{[a_{j+1}}_{b_{j+1}} \dots \delta^{a_n]}_{b_n}. \quad (5.4.8)$$

Proof [Optional Reading]

Here we only give the main steps. The reader should fill in the details of the proof for each step. First, one can show that

$$\delta^{[a_1}_{a_1} \delta^{a_2}_{b_2} \dots \delta^{a_n]}_{b_n} = \frac{1}{n} \delta^{[a_2}_{b_2} \dots \delta^{a_n]}_{b_n},$$

$$\delta^{[a_2}{}_{a_2} \delta^{a_3}{}_{b_3} \cdots \delta^{a_n]}{}_{b_n} = \frac{2}{n-1} \delta^{[a_3}{}_{b_3} \cdots \delta^{a_n]}{}_{b_n},$$

and carrying over to the general case,

$$\delta^{[a_j}{}_{a_j} \delta^{a_{j+1}}{}_{b_{j+1}} \cdots \delta^{a_n]}{}_{b_n} = \frac{j}{n-(j-1)} \delta^{[a_{j+1}}{}_{b_{j+1}} \cdots \delta^{a_n]}{}_{b_n}.$$

Therefore, it can be proved that

$$\begin{aligned} \delta^{[a_1}{}_{a_1} \cdots \delta^{a_j}{}_{a_j} \delta^{a_{j+1}}{}_{b_{j+1}} \cdots \delta^{a_n]}{}_{b_n} &= \frac{1}{n} \frac{2}{n-1} \frac{3}{n-2} \cdots \frac{j}{n-j+1} \delta^{[a_{j+1}}{}_{b_{j+1}} \cdots \delta^{a_n]}{}_{b_n} \\ &= \frac{(n-j)! j!}{n!} \delta^{[a_{j+1}}{}_{b_{j+1}} \cdots \delta^{a_n]}{}_{b_n}. \end{aligned} \quad \square$$

Theorem 5.4.4

$$(a) \varepsilon^{a_1 \cdots a_n} \varepsilon_{b_1 \cdots b_n} = (-1)^s n! \delta^{[a_1}{}_{b_1} \cdots \delta^{a_n]}{}_{b_n}, \quad (5.4.9)$$

$$(b) \varepsilon^{a_1 \cdots a_j a_{j+1} \cdots a_n} \varepsilon_{a_1 \cdots a_j b_{j+1} \cdots b_n} = (-1)^s (n-j)! j! \delta^{[a_{j+1}}{}_{b_{j+1}} \cdots \delta^{a_n]}{}_{b_n}. \quad (5.4.10)$$

Proof $\varepsilon^{a_1 \cdots a_n} \varepsilon_{b_1 \cdots b_n} = \varepsilon^{[a_1 \cdots a_n]} \varepsilon_{[b_1 \cdots b_n]}$ indicates that all the upper indices and all the lower indices of $\varepsilon^{a_1 \cdots a_n} \varepsilon_{b_1 \cdots b_n}$ are antisymmetric. It is not difficult to prove that the collection of all tensors of type (n, n) satisfying this condition is a 1-dimensional vector space, and since $\delta^{[a_1}{}_{b_1} \cdots \delta^{a_n]}{}_{b_n} = \delta^{[a_1}{}_{[b_1} \cdots \delta^{a_n]}{}_{b_n]}$, any tensor in this collection can only differ by a multiplicative factor. Thus, $\varepsilon^{a_1 \cdots a_n} \varepsilon_{b_1 \cdots b_n} = K \delta^{[a_1}{}_{b_1} \cdots \delta^{a_n]}{}_{b_n}$. Contracting with $\varepsilon_{a_1 \cdots a_n} \varepsilon^{b_1 \cdots b_n}$, the left-hand side yields $(-1)^s n! (-1)^s n!$, and the right-hand yields $K \varepsilon_{b_1 \cdots b_n} \varepsilon^{b_1 \cdots b_n} = K (-1)^s n!$, and hence $K = (-1)^s n!$, which brings (5.4.9). Contracting the first j upper and lower indices on both sides gives

$$\begin{aligned} \varepsilon^{a_1 \cdots a_j a_{j+1} \cdots a_n} \varepsilon_{a_1 \cdots a_j b_{j+1} \cdots b_n} &= (-1)^s n! \delta^{[a_1}{}_{a_1} \cdots \delta^{a_j}{}_{a_j} \delta^{a_{j+1}}{}_{b_{j+1}} \cdots \delta^{a_n]}{}_{b_n} \\ &= (-1)^s (n-j)! j! \delta^{[a_{j+1}}{}_{b_{j+1}} \cdots \delta^{a_n]}{}_{b_n}. \end{aligned}$$

(Lemma 5.4.3 is used in the last step.) Thus, we arrive at (5.4.10). \square

5.5 Integrating Functions on Manifolds, Gauss's Theorem

Definition 1 Suppose ε is an arbitrary volume element on a manifold M , and f is a C^0 function on M , then the **integral of f on M** (denoted by $\int_M f$) is defined as the integral of the n -form field $f\varepsilon$ on M , i.e.,

$$\int_M f := \int_M f \varepsilon. \quad (5.5.1)$$

From Definition 1 we see that the integral of a function depends on the choice of a volume element. As long as a metric is given on the manifold, we

stipulate that the integral of a function is always defined using the associated volume element. In this way, for an oriented manifold with a metric, the integral of a given function is determined. Take the 3-dimensional Euclidean space $(\mathbb{R}^3, \delta_{ab})$ as an example. Suppose $\{x, y, z\}$ is a right-handed Cartesian coordinate system, then $\epsilon = dx \wedge dy \wedge dz$ is an associated volume element, and hence the integral of a function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ on $(\mathbb{R}^3, \delta_{ab})$ is, by definition, $\int_{\mathbb{R}^3} f = \int_{\mathbb{R}^3} f \epsilon$. The right-hand side is nothing but an integral of a 3-form field $\omega \equiv f \epsilon$, and according to its definition (Definition 4 of Sect. 5.2), one should express ω in the form of (5.2.1) using the dual basis of the right-handed system. Let $F(x, y, z)$ be the function of 3 variables coming from combining f with the Cartesian system $\{x, y, z\}$, then

$$\omega = F(x, y, z) dx \wedge dy \wedge dz.$$

[This is a special case of (5.2.1).] Hence,

$$\int f = \int f \epsilon = \int \omega = \iiint F(x, y, z) dx \wedge dy \wedge dz.$$

If you like, you can also compute it using the (right-handed) spherical coordinate system $\{r, \theta, \varphi\}$. It follows from the line element $ds^2 = dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2)$ that $g = r^4 \sin^2 \theta$, and thus from (5.4.4) we know that $\epsilon = r^2 \sin \theta dr \wedge d\theta \wedge d\varphi$. Therefore, (5.2.1) in the present case is $\omega \equiv f \epsilon = \hat{F}(r, \theta, \varphi) r^2 \sin \theta dr \wedge d\theta \wedge d\varphi$ [where $\hat{F}(r, \theta, \varphi)$ comes from combining f with $\{r, \theta, \varphi\}$]. Hence,

$$\int f = \int f \epsilon = \int \omega = \iiint \hat{F}(r, \theta, \varphi) r^2 \sin \theta dr \wedge d\theta \wedge d\varphi.$$

Now we will introduce the general form of Gauss's theorem. The form of Gauss's law that is familiar to readers is

$$\iiint_V (\vec{\nabla} \cdot \vec{A}) dV = \oint_S \vec{A} \cdot \vec{n} dS. \quad (5.5.2)$$

Respectively, the two sides of the above equation can be colloquially described as “the integral of the product of the function $\vec{\nabla} \cdot \vec{A}$ and the volume element dV ” and “the integral of the product of the function $\vec{A} \cdot \vec{n}$ and the area element (2-dimensional volume element) dS ”. Now we will show in two steps that the Stokes theorem (5.3.1) leads to a formula which includes (5.5.2) as a special case. The first step is to derive Theorem 5.5.1, the left-hand side of which can be seen as a generalization of (5.5.2).

Theorem 5.5.1 *Suppose M is an n -dimensional oriented manifold, N is an n -dimensional compact embedded submanifold with boundary in M , g_{ab} is a metric on M , ϵ and ∇_a are the associated volume element and the associated derivative operator, and v^a is a C^1 vector field on M , then*

$$\int_{i(N)} (\nabla_b v^b) \epsilon = \int_{\partial N} v^b \epsilon_{ba_1 \dots a_{n-1}}. \quad (5.5.3)$$

Remark 1 The left-hand side of the equation above can be seen as a generalization of the left-hand side of (5.5.2).

Proof The exterior derivative of the $(n - 1)$ -form field $\omega_{a_1 \dots a_{n-1}} \equiv v^b \epsilon_{ba_1 \dots a_{n-1}}$ is the n -form field $(d\omega)_{ca_1 \dots a_{n-1}} = n \nabla_c (v^b \epsilon_{|b|a_1 \dots a_{n-1}})$, in which ∇_c can be any torsion-free derivative operator. The collection of all n -forms at any point in N is a 1-dimensional vector space. Hence, two n -forms $d\omega$ and ϵ only differ by a multiplicative factor, i.e.,

$$n \nabla_c (v^b \epsilon_{|b|a_1 \dots a_{n-1}}) = h \epsilon_{ca_1 \dots a_{n-1}}, \quad (5.5.4)$$

where h is a function on N that can be found as follows: contracting both sides with $\epsilon^{ca_1 \dots a_{n-1}}$ the right-hand side yields $(-1)^s hn!$, and the left-hand side yields

$$\begin{aligned} n \epsilon^{ca_1 \dots a_{n-1}} \nabla_c (v^b \epsilon_{|b|a_1 \dots a_{n-1}}) &= n \epsilon^{[ca_1 \dots a_{n-1}]} \nabla_c (v^b \epsilon_{ba_1 \dots a_{n-1}}) \\ &= n \epsilon^{ca_1 \dots a_{n-1}} \epsilon_{ba_1 \dots a_{n-1}} \nabla_c v^b = n(-1)^s (n-1)! \delta^c_b \nabla_c v^b = (-1)^s n! \nabla_b v^b. \end{aligned}$$

[Theorem 2.6.2(a) is used in the first equality; we stipulate ∇_c to be associated with g_{ab} starting from the second step; in the third equality we used (5.4.10).] Hence, $h = \nabla_b v^b$, and $d\omega = \epsilon \nabla_b v^b$. Therefore, the Stokes theorem in this case takes the form of (5.5.3). \square

Now we go one step further and rewrite the right-hand side of (5.5.3) into a form like the right-hand side of (5.5.2). Since the latter involves the volume element dS on the boundary S , let us start with the volume element of ∂N . Here we only talk about the case where ∂N is not a null hypersurface, and thus we can talk about the normalized normal vector n^a of ∂N that satisfies $n^a n_a = \pm 1$ (see Sect. 4.4). The induced metric of the metric g_{ab} on ∂N is $h_{ab} = g_{ab} \mp n_a n_b$ [see (4.4.2)]. Regarding ∂N as an $(n - 1)$ -dimensional manifold with the metric h_{ab} , its volume element (denoted by $\hat{\epsilon}_{a_1 \dots a_{n-1}}$) should satisfy two conditions: ① compatible with the induced orientation of ∂N (denoted by $\bar{\epsilon}_{a_1 \dots a_{n-1}}$, see Remark 1 of Sect. 5.3); ② associated with h_{ab} , i.e.,

$$\hat{\epsilon}^{a_1 \dots a_{n-1}} \hat{\epsilon}_{a_1 \dots a_{n-1}} = (-1)^{\hat{s}} (n-1)!, \quad (5.5.5)$$

where $\hat{\epsilon}^{a_1 \dots a_{n-1}}$ is the result of raising the indices of $\hat{\epsilon}_{a_1 \dots a_{n-1}}$ using h_{ab} , and \hat{s} is the number of negative numbers in the diagonal elements of h_{ab} . The volume element $\hat{\epsilon}_{a_1 \dots a_{n-1}}$ on ∂N that satisfies these two conditions is called the **induced volume element**. Suppose n^b is the outgoing unit normal vector of ∂N [with $i(N)$ being the interior, there is a clear meaning for “outgoing”], then the induced volume element $\hat{\epsilon}_{a_1 \dots a_{n-1}}$ and the volume element $\epsilon_{ba_1 \dots a_{n-1}}$ on N have the following relation (for a proof, see Optional Reading 5.5.1):

$$\hat{\epsilon}_{a_1 \dots a_{n-1}} = n^b \epsilon_{ba_1 \dots a_{n-1}}. \quad (5.5.6)$$

[Optional Reading 5.5.1]

Now we will show that the $\hat{\varepsilon}_{a_1 \dots a_{n-1}}$ in the equation above does satisfy the two conditions of the induced volume element. $\forall q \in \partial N$, suppose $\{(e_\mu)^a\}$ is the right-handed orthonormal basis at q satisfying $(e_1)^a = n^a$, then

$$\varepsilon_{a_1 \dots a_n} = (e^1 \wedge \dots \wedge e^n)_{a_1 \dots a_n} = \pm n_{a_1} \wedge (e^2 \wedge \dots \wedge e^n)_{a_2 \dots a_n}.$$

From the spirit of Remark 1 in Sect. 5.3 [see, Wald (1984) p. 431 for details] we know that $(e^2 \wedge \dots \wedge e^n)_{a_2 \dots a_n}$ serves as the induced orientation $\bar{\varepsilon}_{a_2 \dots a_n}$ at $q \in \partial N$, and hence

$$\varepsilon_{a_1 \dots a_n} = \pm n_{a_1} \wedge \bar{\varepsilon}_{a_2 \dots a_n}, \text{ also written as } \varepsilon_{ba_1 \dots a_{n-1}} = \pm n_b \wedge \bar{\varepsilon}_{a_1 \dots a_{n-1}}.$$

Using this, one can easily show that $\bar{\varepsilon}_{a_1 \dots a_{n-1}} = n^b \varepsilon_{ba_1 \dots a_{n-1}}$, and then it follows from (5.5.6) that $\hat{\varepsilon}_{a_1 \dots a_{n-1}} = +1 \cdot \bar{\varepsilon}_{a_1 \dots a_{n-1}}$. Thus, $\hat{\varepsilon}_{a_1 \dots a_{n-1}}$ is compatible with the induced orientation $\bar{\varepsilon}_{a_1 \dots a_{n-1}}$, i.e., condition ① is satisfied. As an exercise (Exercise 5.10), the reader should verify that $\hat{\varepsilon}_{a_1 \dots a_{n-1}} = n^b \varepsilon_{ba_1 \dots a_{n-1}}$ also satisfies condition ②, i.e., (5.5.5). Note that condition ② can only determine $\hat{\varepsilon}_{a_1 \dots a_{n-1}}$ up to a minus sign [i.e., $\hat{\varepsilon}_{a_1 \dots a_{n-1}} = -n^b \varepsilon_{ba_1 \dots a_{n-1}}$ also satisfies (5.5.5)]. Only when taken together with condition ① can $\hat{\varepsilon}_{a_1 \dots a_{n-1}}$ be determined as $n^b \varepsilon_{ba_1 \dots a_{n-1}}$.

[The End of Optional Reading 5.5.1]

The theorem below is the general version of Gauss's theorem that contains (5.5.2) as a special case.

Theorem 5.5.2 (Gauss's Theorem) Suppose M is an n -dimensional oriented manifold, N is an n -dimensional compact submanifold with boundary in M , g_{ab} is a metric on M , ϵ and ∇_a are, respectively, the associated volume element and the associated derivative operator, $\hat{\epsilon}$ is the induced volume element on ∂N , n^a is the outgoing normal vector of ∂N satisfying $n^a n_a = \pm 1$, and v^a is a C^1 vector field on M . Then,

$$\int_{i(N)} (\nabla_a v^a) \epsilon = \pm \int_{\partial N} v^a n_a \hat{\epsilon}. \quad (+\text{for } n^a n_a = +1, -\text{for } n^a n_a = -1.) \quad (5.5.7)$$

Proof From Theorem 5.5.1 we know that all we have to prove is $\int_{\partial N} v^b \varepsilon_{ba_1 \dots a_{n-1}} = \pm \int_{\partial N} v^a n_a \hat{\epsilon}$. Let $\omega_{a_1 \dots a_{n-1}} = v^b \varepsilon_{ba_1 \dots a_{n-1}}$. Noticing the discussion at the end of Sect. 5.2 about $\int_{\phi[S]} \omega \equiv \int_{\phi[S]} \tilde{\omega}$, we can see that here $\int_{\partial N} v^b \varepsilon_{ba_1 \dots a_{n-1}}$ is $\int_{\partial N} \tilde{\omega}$. Hence, all we have to prove is that

$$\tilde{\omega}_{a_1 \dots a_{n-1}} = \pm v^b n_b \hat{\epsilon}_{a_1 \dots a_{n-1}}, \quad \forall q \in \partial N, \quad (5.5.8)$$

where n^a is the outgoing unit normal vector of ∂N . Both sides of the above equation are $(n-1)$ -forms on W_q , and hence there exists a K such that

$$\tilde{\omega}_{a_1 \dots a_{n-1}} = K v^b n_b \hat{\epsilon}_{a_1 \dots a_{n-1}}, \quad (5.5.9)$$

and thus all we have to prove is that $K = \pm 1$. Suppose $\{(e_0)^a = n^a, (e_1)^a, \dots, (e_{n-1})^a\}$ is a right-handed orthonormal basis of V_q . Contracting $(e_1)^{a_1} \dots (e_{n-1})^{a_{n-1}}$

with the equation above, the right-hand side gives

$$K v^b n_b \hat{\varepsilon}_{12\cdots(n-1)} = \pm K v^b (e^0)_b \hat{\varepsilon}_{12\cdots(n-1)} = \pm K v^0, \quad (5.5.10)$$

where we used $n_b = \pm (e^0)_b$ in the first equality; in the second equality we used the following fact: it can be shown from the definition of the induced orientation $\bar{\varepsilon}$ that the right-handedness of $\{(e_0)^a = n^a, (e_1)^a, \dots, (e_{n-1})^a\}$ (measured by the orientation ε) assures the right-handedness of $\{(e_1)^a, \dots, (e_{n-1})^a\}$ (measured by $\bar{\varepsilon}$), and thus $\hat{\varepsilon}_{12\cdots(n-1)} = 1$. On the other hand, the left-hand side of (5.5.9) after the contraction yields

$$\begin{aligned} \tilde{\omega}_{a_1\cdots a_{n-1}}(e_1)^{a_1} \cdots (e_{n-1})^{a_{n-1}} &= \omega_{a_1\cdots a_{n-1}}(e_1)^{a_1} \cdots (e_{n-1})^{a_{n-1}} \\ &= v^b \varepsilon_{ba_1\cdots a_{n-1}}(e_1)^{a_1} \cdots (e_{n-1})^{a_{n-1}} = v^\mu \varepsilon_{\mu 12\cdots(n-1)} = v^0 \varepsilon_{012\cdots(n-1)} = v^0, \end{aligned} \quad (5.5.11)$$

where (5.2.7) is used in the first equality, and the right-handedness of $\{(e_0)^a = n^a, (e_1)^a, \dots, (e_{n-1})^a\}$ is used in the fifth equality. Comparing (5.5.10) and (5.5.11) we obtain $K = \pm 1$. \square

Remark 2 One of the conditions for (5.5.7) is that n^a is the outgoing unit normal vector of ∂N . If we change the stipulation to “ n^a is outgoing when $n^a n_a = +1$, n^a is ingoing [pointing towards $i(N)$] when $n^a n_a = -1$ ”, then the \pm sign in the right-hand side of (5.5.7) vanishes, and Gauss’s theorem turns into the following form

$$\int_{i(N)} (\nabla_a v^a) \varepsilon = \int_{\partial N} v^a n_a \hat{\varepsilon}. \quad (5.5.7')$$

If ∂N is a null hypersurface, i.e., $n^a n_a = 0$, then (5.5.7') still holds; however, $\hat{\varepsilon}$ needs to be defined as follows (sans proof):

$$\frac{1}{n} \varepsilon_{a_1\cdots a_n} = n_{[a_1} \hat{\varepsilon}_{a_2\cdots a_n]}.$$

5.6 Dual Differential Forms

Use $\Lambda_p(l)$ to represent the collection of all l -forms ($l \leq n$) at $p \in M$. It follows from (5.1.4) that

$$\dim \Lambda_p(l) = \frac{n!}{l!(n-l)!} = \dim \Lambda_p(n-l).$$

If M is an oriented manifold with a metric g_{ab} and ε is the associated volume element, then we can define an isomorphism between $\Lambda_M(l)$ and $\Lambda_M(n-l)$ using ε and g_{ab} as follows:

Definition 1 $\forall \omega \in \Lambda_M(l)$, define its **dual form** ${}^*\omega \in \Lambda_M(n-l)$ as

$${}^*\omega_{a_1 \dots a_{n-l}} := \frac{1}{l!} \omega^{b_1 \dots b_l} \varepsilon_{b_1 \dots b_l a_1 \dots a_{n-l}}, \quad (5.6.1)$$

where

$$\omega^{b_1 \dots b_l} = g^{b_1 c_1} \dots g^{b_l c_l} \omega_{c_1 \dots c_l}.$$

Remark 1 The $*$ operator we defined above is called the **Hodge star**, and ${}^*\omega$ is also called the **Hodge dual** of the form ω . It is not difficult to see that: ① $* : \Lambda_M(l) \rightarrow \Lambda_M(n-l)$ is an isomorphism; ② for a 0-form field $f \in \mathcal{F}_M$, its dual form field by definition is

$${}^*f_{a_1 \dots a_n} = \frac{1}{0!} f \varepsilon_{a_1 \dots a_n} = f \varepsilon_{a_1 \dots a_n},$$

i.e., *f equals f times the volume element ε associated with the metric. Therefore, one can say that the integral of a function f is defined as the integral of its dual form field. Applying $*$ to the above equation again we have

$${}^*({}^*f) = {}^*(f \varepsilon) = \frac{1}{n!} f \varepsilon^{b_1 \dots b_n} \varepsilon_{b_1 \dots b_n} = (-1)^s f.$$

[Equation (5.4.3) is used in the third equality.] This result can be generalized into the following theorem:

Theorem 5.6.1

$${}^{**}\omega = (-1)^{s+l(n-l)} \omega. \quad (5.6.2)$$

Proof Exercise 5.11. □

Now, from the differential geometry point of view, let us revisit the vector algebra and vector field theory on 3-dimensional Euclidean space $(\mathbb{R}^3, \delta_{ab})$ that we are already familiar with (where M is \mathbb{R}^3).

(1) Why have we never heard of 1-, 2- and 3-form fields before? First, using the Euclidean metric δ_{ab} , one can turn a dual vector field ω_a into a vector field $\omega^a = \delta^{ab} \omega_b$, which eliminates the need to use a 1-form field. Later on, we will not distinguish the upper and lower indices strictly when we are dealing only with $(\mathbb{R}^3, \delta_{ab})$. Second, since $n = 3$, $\Lambda_M(2)$ and $\Lambda_M(1)$ have the same dimension, and $\omega \in \Lambda_M(2)$ and ${}^*\omega \in \Lambda_M(1)$ can be identified using the isomorphism $* : \Lambda_M(2) \rightarrow \Lambda_M(1)$, which eliminates the need to use a 2-form field. Similarly, $\Lambda_M(3)$ and $\Lambda_M(0)$ have the same dimension, and using the isomorphism $* : \Lambda_M(3) \rightarrow \Lambda_M(0)$ one can identify $\omega \in \Lambda_M(3)$ and ${}^*\omega \in \Lambda_M(0)$, the latter of which is exactly a function on \mathbb{R}^3 (a 0-form field). Therefore, any differential form on the 3-dimensional Euclidean space can be represented by a function or a vector field.

(2) Now we discuss the dot product and cross product operations of the vector algebra. Denote the vectors \vec{A} and \vec{B} as A^a and B^a , respectively. Naturally, the dot

product of \vec{A} and \vec{B} will be $A_a B^a$. However, how should we understand the cross product $\vec{A} \times \vec{B}$? Let

$$\omega_{ab} \equiv A_a \wedge B_b = 2A_{[a} B_{b]} , \quad (\text{where } A_a \equiv \delta_{ab} A^b, B_b \equiv \delta_{ba} B^a)$$

then

$${}^*\omega_c = \frac{1}{2} \omega^{ab} \varepsilon_{abc} = \varepsilon_{abc} A^{[a} B^{b]} = \varepsilon_{abc} A^a B^b , \quad (5.6.3)$$

where ε_{abc} is the volume element associated with the Euclidean metric. Suppose $\{x, y, z\}$ is a right-handed Cartesian coordinate system, then its coordinate basis is orthonormal. It follows from (5.4.2) that the nonzero components ε_{ijk} of ε_{abc} in this system are

$$\varepsilon_{123} = \varepsilon_{312} = \varepsilon_{231} = -\varepsilon_{132} = -\varepsilon_{321} = -\varepsilon_{213} = 1 ,$$

and thus ε_{ijk} is the familiar Levi-Civita symbol. Therefore, the k th component of ${}^*\omega_c$ in this Cartesian system is

$${}^*\omega_k = \varepsilon_{ijk} A^i B^j , \quad k = 1, 2, 3 . \quad (5.6.4)$$

According to the definition of $\vec{A} \times \vec{B}$, the right-hand side of the above equation is exactly the k th component $(\vec{A} \times \vec{B})_k$. Hence, $\vec{A} \times \vec{B}$ can be viewed as ${}^*\omega$ (or more precisely, the vector corresponding to the dual vector ${}^*\omega$). Also $\omega = \mathbf{A} \wedge \mathbf{B}$, and thus finding the cross product of \vec{A} and \vec{B} is the same as finding the wedge product $\mathbf{A} \wedge \mathbf{B}$ and then taking its dual. This can be expressed simply as $\times = {}^* \circ \wedge$.

(3) Now let us look at the vector field theory of the 3-dimensional Euclidean space from the viewpoint of differential geometry. As we mentioned previously, $\vec{\nabla}$ in the vector field theory is the derivative operator ∂_a associated with the Euclidean metric δ_{ab} ; in principle, any equation that involves $\vec{\nabla}$ can be expressed in terms of ∂_a . For instance:

- (a) $\vec{\nabla} f = \partial_a f$;
 - (b) $\vec{\nabla} \cdot \vec{A} = \partial_a A^a$;
 - (c) $\vec{\nabla} \times \vec{A} = \varepsilon^{abc} \partial_a A_b$ [the derivation is similar to (5.6.3)];
 - (d) $\vec{\nabla} \cdot (\vec{A} \vec{B}) = \partial_a (A^a B^b)$;
 - (e) $\vec{\nabla} \vec{A} = \partial^a A^b$;
 - (f) $\nabla^2 f = \partial_a \partial^a f$;
 - (g) $\nabla^2 \vec{A} = \partial_a \partial^a A^b$.
- (5.6.5)

By means of ∂_a and the abstract index notation, one can also simplify the derivation of some useful formulas and make the reasoning clearer. Here we give only two examples.

Example 1 Using ∂_a , show that

$$\vec{\nabla} \cdot (\vec{A} \times \vec{B}) = \vec{B} \cdot (\vec{\nabla} \times \vec{A}) - \vec{A} \cdot (\vec{\nabla} \times \vec{B}). \quad (5.6.6)$$

Proof

$$\vec{\nabla} \cdot (\vec{A} \times \vec{B}) = \partial_c (\varepsilon^{cab} A_a B_b) = \varepsilon^{cab} (A_a \partial_c B_b + B_b \partial_c A_a), \quad (5.6.7)$$

while

$$\begin{aligned} \vec{B} \cdot (\vec{\nabla} \times \vec{A}) &= B_b (\vec{\nabla} \times \vec{A})^b = B_b \varepsilon^{bca} \partial_c A_a = \varepsilon^{cab} B_b \partial_c A_a, \\ -\vec{A} \cdot (\vec{\nabla} \times \vec{B}) &= -A_a (\vec{\nabla} \times \vec{B})^a = -A_a \varepsilon^{acb} \partial_c B_b = \varepsilon^{cab} A_a \partial_c B_b. \end{aligned}$$

Plugging into (5.6.7) we get (5.6.6). \square

Example 2 Using ∂_a , show that

$$\vec{\nabla}(\vec{A} \cdot \vec{B}) = (\vec{A} \cdot \vec{\nabla})\vec{B} + (\vec{B} \cdot \vec{\nabla})\vec{A} + \vec{A} \times (\vec{\nabla} \times \vec{B}) + \vec{B} \times (\vec{\nabla} \times \vec{A}). \quad (5.6.8)$$

Proof For each term on the right-hand side of (5.6.8), we have

$$\text{the first term} = A_a \partial^a B^b, \quad \text{the second term} = B_a \partial^a A^b,$$

$$\begin{aligned} \text{the third term} &= \vec{A} \times (\varepsilon^{cde} \partial_d B_e) = \varepsilon^{bac} A_a (\varepsilon_{cde} \partial^d B^e) \\ &= 2\delta_{[d}^b \delta_{e]}^a A_a \partial^d B^e = (\delta_d^b \delta_e^a - \delta_e^b \delta_d^a) A_a \partial^d B^e = A_a \partial^b B^a - A_a \partial^a B^b, \end{aligned}$$

and similarly,

$$\text{the fourth term} = B_a \partial^b A^a - B_a \partial^a A^b.$$

Hence,

$$\text{the r.h.s. of (5.6.8)} = A_a \partial^b B^a + B_a \partial^b A^a = \partial^b (A_a B^a) = \vec{\nabla}(\vec{A} \cdot \vec{B}).$$

\square

(4) The gradient, curl and divergence in the 3-dimensional Euclidean space can be simply expressed using the exterior differentiation as follows:

Theorem 5.6.2 Suppose f and \vec{A} are respectively a function and a vector field on the 3-dimensional Euclidean space, then

$$\text{grad } f = df, \quad \text{curl } \vec{A} = {}^*d\vec{A}, \quad \text{div } \vec{A} = {}^*d({}^*\vec{A}). \quad (5.6.9)$$

Proof Exercise 5.11. \square

The fact that \mathbb{R}^3 is a trivial manifold assures that a closed form field on \mathbb{R}^3 is exact (see Remark 1 of Sect. 5.1). Combining this with (5.6.9), one can easily prove (Exercise 5.15) the following well-known propositions which are not so straightforward to prove by the standard vector analysis of 3-dimensional Euclidean space:

(1) a vector with no curl must be a gradient field, i.e.,

$$\operatorname{curl} \vec{E} = 0 \Rightarrow \exists \text{ a scalar field } \phi \text{ s.t. } \vec{E} = \operatorname{grad} \phi,$$

(2) a vector with no divergence must be a curl field, i.e.,

$$\operatorname{div} \vec{B} = 0 \Rightarrow \exists \text{ a vector field } \vec{A} \text{ s.t. } \vec{B} = \operatorname{curl} \vec{A}.$$

5.7 Computing the Riemann Curvature Using the Tetrad Method [Optional Reading]

There are two major methods for computing the Riemann curvature $R_{abc}{}^d$ of a derivative operator ∇_a . The first one uses a coordinate basis field; the second one uses a non-coordinate basis field. In Sect. 3.4.2 we have already introduced the first method, in which the key step is to find the manifestation of ∇_a in a coordinate basis field, namely the Christoffel symbol $\Gamma^\sigma{}_{\mu\tau}$. This section will discuss how to compute $R_{abc}{}^d$ using a non-coordinate basis field. First, we need to find the manifestation of ∇_a in this non-coordinate basis field. For a given derivative operator ∇_a , suppose $\{(e_\mu)^a\}$ is an arbitrary basis field whose domain is $U \subset M$. The derivative of the μ th basis field $(e_\mu)^a$ along the τ th basis field $(e_\tau)^a$, i.e., $(e_\tau)^b \nabla_b (e_\mu)^a$, is also a vector field on U , and thus can be expanded in terms of the basis field $\{(e_\sigma)^a\}$:

$$(e_\tau)^b \nabla_b (e_\mu)^a = \gamma^\sigma{}_{\mu\tau} (e_\sigma)^a, \quad (5.7.1)$$

where $\gamma^\sigma{}_{\mu\tau}$, called the **connection coefficients**, can be regarded as the manifestation of ∇_a in the basis field $\{(e_\sigma)^a\}$. The $\gamma^\sigma{}_{\mu\tau}$ of a coordinate basis field are specifically denoted by $\Gamma^\sigma{}_{\mu\tau}$. It is not difficult to show that (Exercise 5.17) these $\Gamma^\sigma{}_{\mu\tau}$ are exactly the components of the Christoffel symbol $\Gamma^c{}_{ab}$ defined in Sect. 3.1 in this coordinate basis field. That is, the coordinate components of a Christoffel symbol can be defined equivalently as follows:

$$\left(\frac{\partial}{\partial x^\tau} \right)^b \nabla_b \left(\frac{\partial}{\partial x^\mu} \right)^a = \Gamma^\sigma{}_{\mu\tau} \left(\frac{\partial}{\partial x^\sigma} \right)^a. \quad (5.7.2)$$

The contraction of (5.7.1) and the dual basis $(e^\nu)_a$ gives the explicit expression for $\gamma^\sigma{}_{\mu\tau}$:

$$\gamma^\nu{}_{\mu\tau} = (e^\nu)_a (e_\tau)^b \nabla_b (e_\mu)^a. \quad (5.7.3)$$

τ can be chosen from 1 to n for given values of μ and ν . Thus, $\{\gamma^\nu{}_{\mu\tau} | \nu, \mu$ are fixed, $\tau = 1, \dots, n\}$ is the collection of n real functions $\gamma^\nu{}_{\mu 1}, \dots, \gamma^\nu{}_{\mu n}$. Using these components we can define a 1-form $(\omega_\mu{}^\nu)_a$, called the **connection 1-form** of ∇_a in the basis field $\{(e_\mu)^a\}$, denoted by $\omega_\mu{}^\nu{}_a$ for short, as follows:

$$\omega_\mu{}^\nu{}_a := -\gamma^\nu{}_{\mu\tau} (e^\tau)_a. \quad (5.7.4)$$

Note that the lower index a of $\omega_{\mu}{}^v{}_a$ is an abstract index, indicating it is a 1-form; μ and v are the indices numbering the connection 1-forms. It is easy to derive from the equation above and (5.7.3) that

$$\omega_{\mu}{}^v{}_a = -(e^v)_c \nabla_a (e_\mu)^c = (e_\mu)^c \nabla_a (e^v)_c, \quad (5.7.5)$$

where in the first equality we used $(e^\tau)_a (e_\tau)^b = \delta^b{}_a$ [see (2.6.4)], and in the second equality we used the Leibniz rule and the definition of a dual basis $(e^v)_c (e_\mu)^c = \delta^v{}_\mu$. The collection of all the connection 1-forms $\{\omega_{\mu}{}^v{}_a | \mu, v = 1, \dots, n\}$ can be viewed as the manifestation of ∇_a in the basis field $\{(e_\mu)^a\}$. For a given ∇_a , in principle one can choose a basis field $\{(e_\mu)^a\}$ and compute all the connection 1-forms $\omega_{\mu}{}^v{}_a$ of ∇_a with respect to this basis field, and then compute the curvature tensor. A basis is also called a **frame** (a 4-dimensional frame is also called a **tetrad**). In many cases, a frame actually will mean a non-coordinate basis. Now, we will present the tetrad method for computing the curvature introduced by Élie Cartan.

Since both $\omega_{\mu}{}^v{}_a$ and the dual basis $(e^\mu)_c$ are 1-forms, we can drop the lower index a and denote them as $\omega_\mu{}^v$ and e^μ , respectively. Under the torsion-free condition, they have the following relation:

Theorem 5.7.1 (Cartan's first equation of structure)

$$de^v = -e^\mu \wedge \omega_\mu{}^v. \quad (5.7.6)$$

Proof

$$\begin{aligned} -(e^\mu)_a \wedge \omega_\mu{}^v{}_b &= -(e^\mu)_a \wedge [(e_\mu)^c \nabla_b (e^v)_c] = -2(e^\mu)_{[a} (e_\mu)^c \nabla_{b]} (e^v)_c \\ &= -2\delta^c{}_{[a} \nabla_{b]} (e^v)_c = -2\nabla_{[b} (e^v)_{a]} = (de^v)_{ab}. \end{aligned} \quad \square$$

Now we discuss how to calculate the curvature tensor $R_{abc}{}^d$ from $\omega_\mu{}^v$. Let

$$R_{ab\mu}{}^v \equiv R_{abc}{}^d (e_\mu)^c (e^v)_d. \quad (5.7.7)$$

Then $R_{ab\mu}{}^v = -R_{ba\mu}{}^v$ indicates that $R_{ab\mu}{}^v$ can be regarded as the μ th and v th 2-form fields, denoted for short as $R_\mu{}^v$. It follows from (5.7.7) that μ and v are the component indices (for the frame components) of $R_{abc}{}^d$, while they can also be viewed as the indices numbering the 2-form fields $R_\mu{}^v$ (however, the μ and v in $\omega_\mu{}^v$ can only be viewed as the indices numbering the 1-forms). The curvature 2-forms $R_\mu{}^v$ and the connection 1-forms $\omega_\mu{}^v$ have the following relation:

Theorem 5.7.2 (Cartan's second equation of structure)

$$R_\mu{}^v = d\omega_\mu{}^v + \omega_\mu{}^\lambda \wedge \omega_\lambda{}^v. \quad (5.7.8)$$

Proof It follows from (5.7.7) and the definition of $R_{abc}{}^d$ that

$$R_{ab\mu}{}^v = 2(e_\mu)^c \nabla_{[a} (e^v)_{b]}.$$

Also,

$$\begin{aligned}
(e_\mu)^c \nabla_a \nabla_b (e^\nu)_c &= \nabla_a [(e_\mu)^c \nabla_b (e^\nu)_c] - [\nabla_a (e_\mu)^c] \nabla_b (e^\nu)_c \\
&= \nabla_a \omega_\mu^v b - [\nabla_a (e_\mu)^d] \delta^c_d \nabla_b (e^\nu)_c \\
&= \nabla_a \omega_\mu^v b - [\nabla_a (e_\mu)^d] (e^\lambda)_d (e_\lambda)^c \nabla_b (e^\nu)_c \\
&= \nabla_a \omega_\mu^v b + \omega_\mu^\lambda \omega_\lambda^v b.
\end{aligned}$$

Hence,

$$R_{ab\mu}^v = 2\nabla_{[a} \omega_{\mu}^v{}_{b]} + 2\omega_{\mu}^{\lambda} [\omega_{\lambda}^v{}_{b}] = (\mathrm{d}\omega_{\mu}^v)_{ab} + (\omega_{\mu}^{\lambda} \wedge \omega_{\lambda}^v)_{ab}, \quad (5.7.8')$$

which is exactly (5.7.8). \square

Remark 1 As we mentioned, (5.7.6) only holds for torsion-free connections. When torsion exists, one should add an additional torsion term, and the complete first equation of structure can be written as (also see Appendix I in Volume III):

$$\mathbf{T}^v = \mathrm{d}\mathbf{e}^v + \mathbf{e}^\mu \wedge \boldsymbol{\omega}_\mu^v,$$

where \mathbf{T}^v is the torsion 2-form, which relates to the torsion tensor defined in Exercise 3.1 as $T^v{}_{ab} \equiv T^c{}_{ab} (e^\nu)_c$. Note that the definition of the exterior differentiation (5.1.11) holds only without torsion, so the last step in the proof of Theorem 5.7.1 will not hold in this case.

Remark 2 Equation (5.7.8) is equivalent to (3.4.20'); they are the component expressions for the definition of the curvature (namely the relation between the connection and the curvature) in a frame and in a coordinate basis, respectively.

When $\boldsymbol{\omega}_\mu^v$ are already obtained, we can conveniently derive \mathbf{R}_μ^v using the second equation of structure; all we have to do is to take the exterior differentiation and take the wedge product of $\boldsymbol{\omega}_\mu^v$. To find all the components $R_{\rho\sigma\mu}^v$ of R_{abc}^d in the chosen frame, all we have to do is to take the contraction using the following formula:

$$R_{\rho\sigma\mu}^v = R_{ab\mu}^v (e_\rho)^a (e_\sigma)^b. \quad (5.7.9)$$

Many other works use Ω_i^j (or Θ_i^j), R_{mni}^j and θ^i to denote \mathbf{R}_μ^v , $R_{\rho\sigma\mu}^v$ and \mathbf{e}^μ in this text (note that their i , j and a , b are all component indices), and write the relation between the curvature 2-form Ω_i^j and the tetrad components R_{mni}^j of the curvature tensor as

$$\Omega_i^j = \frac{1}{2} R_{mni}^j \theta^m \wedge \theta^n,$$

Using our notation, this equation may be expressed as

$$\mathbf{R}_\mu^v = \frac{1}{2} R_{\rho\sigma\mu}^v \mathbf{e}^\rho \wedge \mathbf{e}^\sigma, \quad \text{i.e., } R_{ab\mu}^v = \frac{1}{2} R_{\rho\sigma\mu}^v (e^\rho)_a \wedge (e^\sigma)_b. \quad (5.7.10)$$

In fact, this is nothing but a special case of (5.1.6') when $l = 2$.

If, besides ∇_a , there is also a metric g_{ab} given on M , which satisfies $\nabla_a g_{bc} = 0$, then we will have even more to discuss. Using $g_{\mu\nu}$ and $g^{\mu\nu}$ to represent the components of g_{ab} and g^{ab} in the chosen frame, i.e.,

$$g_{\mu\nu} = g_{ab} (e_\mu)^a (e_\nu)^b, \quad (5.7.11)$$

$$g^{\mu\nu} = g^{ab} (e^\mu)_a (e^\nu)_b, \quad (5.7.12)$$

and introducing these two notations:

$$(a) (e_\mu)_a \equiv g_{ab}(e_\mu)^b, \quad (b) (e^\mu)^a \equiv g^{ab}(e^\mu)_b, \quad (5.7.13)$$

we have

$$(a) (e^\mu)_a = g^{\mu\nu}(e_\nu)_a, \quad (b) (e_\mu)^a = g_{\mu\nu}(e^\nu)^a. \quad (5.7.14)$$

To prove (a) of (5.7.14), all we have to do is to verify that both sides acting on $(e_\sigma)^a$ give the same result. To prove (b) of (5.7.14), all we have to do is to verify that both sides acting on $(e^\sigma)_a$ give the same result. The proof is left to the reader.

Raising and lowering the indices for the two equations in (5.7.14) using g^{ab} and g_{ab} yields

$$(a) (e^\mu)^a = g^{\mu\nu}(e_\nu)^a, \quad (b) (e_\mu)_a = g_{\mu\nu}(e^\nu)_a. \quad (5.7.15)$$

In (5.7.14) and (5.7.15), both (a) indicate that the number index ν of the basis can be raised by $g^{\mu\nu}$, and both (b) indicate that the number index ν of the basis can be lowered by $g_{\mu\nu}$. Similarly, one can use $g_{\mu\nu}$ to lower the indices for $\omega_\mu{}^\nu{}_a$, i.e., define (NB: $\omega_{\mu\nu a}$ was meaningless without this definition)

$$\omega_{\mu\nu a} := g_{\nu\sigma}\omega_\mu{}^\sigma{}_a = g_{\nu\sigma}(e_\mu)^c\nabla_a(e^\sigma)_c. \quad (5.7.16)$$

A frame with $g_{\mu\nu}$ as constants (i.e., $\nabla_a g_{\mu\nu} = 0$) is called a **rigid frame**. An orthonormal frame is the simplest rigid frame. For a Lorentzian metric, an orthonormal frame satisfies $g_{\mu\nu} = \eta_{\mu\nu}$, which brings a huge convenience to the calculations (for details, see the example at the end of this chapter). There is another kind of rigid frame that is frequently used in general relativity—the complex null frame, which will be discussed in detail in Sects. 8.7 and 8.8. It is easy to see from (5.7.16) and $\nabla_a g_{\mu\nu} = 0$ that the following relation holds for rigid frames:

$$\omega_{\mu\nu a} = (e_\mu)_b\nabla_a(e_\nu)^b. \quad (5.7.17)$$

Theorem 5.7.3 *For a rigid frame, we have*

$$\omega_{\mu\nu a} = -\omega_{\nu\mu a}. \quad (5.7.18)$$

Proof It follows from (5.7.17) that

$$\begin{aligned} \omega_{\mu\nu a} &= \nabla_a[(e_\mu)_b(e_\nu)^b] - (e_\nu)^b\nabla_a(e_\mu)_b = \nabla_a[g_{bc}(e_\mu)^c(e_\nu)^b] - (e_\nu)^b\nabla_a(e_\mu)_b \\ &= \nabla_a g_{\nu\mu} - (e_\nu)^b\nabla_a(e_\mu)_b = -(e_\nu)^b\nabla_a(e_\mu)_b = -\omega_{\nu\mu a}, \end{aligned}$$

where the fourth equality comes from the fact that $\nabla_a g_{\nu\mu} = 0$. \square

Equation (5.7.18) indicates that, for a rigid frame, the $\omega_{\mu\nu a}$ are antisymmetric with respect to μ and ν , which reduces the number of the independent connection 1-forms from n^2 (where n is the dimension of M) to $n(n-1)/2$ (there are 6 of them when $n=4$). In a chosen basis, the components $\omega_{\mu\nu\rho} \equiv \omega_{\mu\nu a}(e_\rho)^a$ play a similar role in the computation as the Christoffel symbols $\Gamma^\sigma{}_{\mu\tau}$ in a coordinate basis, the former of which also have n^3 numbers, but with only $n^2(n-1)/2$ independent ones (there are 24 of them when $n=4$). Hence, the independent $\omega_{\mu\nu\rho}$ are less than the independent $\Gamma^\sigma{}_{\mu\tau}$. [It follows from the symmetry of its lower indices that there are $n^2(n+1)/2$ independent $\Gamma^\sigma{}_{\mu\tau}$.] $\omega_{\mu\nu\rho}$ are called the **Ricci rotation coefficients**.

The “tetrad method” of computing the curvature tensor includes the following three steps: (a) choosing a tetrad; (b) computing all the connection 1-forms $\omega_\mu{}^\nu$; (c) using Cartan’s second equation of structure (5.7.8) to compute all the curvature 2-forms $R_\mu{}^\nu$ from $\omega_\mu{}^\nu$. Among them, step (b) needs to be further elaborated. Since rigid tetrads are the most commonly

used, here we only introduce the method of computing $\omega_\mu{}^\nu$ using a rigid tetrad. Choosing an arbitrary coordinate system $\{x^\mu\}$, in which we define

$$\Lambda_{\mu\nu\rho} \equiv [(e_v)_{\lambda,\tau} - (e_v)_{\tau,\lambda}](e_\mu)^\lambda(e_\rho)^\tau, \quad (5.7.19)$$

where $(e_v)_\lambda$ and $(e_\mu)^\lambda$ are the λ th components of $(e_v)_a$ and $(e_\mu)^a$, i.e.,

$$(e_v)_\lambda \equiv (e_v)_a(\partial/\partial x^\lambda)^a, \quad (e_\mu)^\lambda \equiv (e_\mu)^a(dx^\lambda)_a,$$

and $(e_v)_{\lambda,\tau}$ is an abbreviation for $\partial(e_v)_\lambda/\partial x^\tau$. It can be easily seen that $\Lambda_{\mu\nu\rho} = -\Lambda_{\rho\nu\mu}$; hence, there are only $n^2(n-1)/2$ independent $\Lambda_{\mu\nu\rho}$. After obtaining all the $\Lambda_{\mu\nu\rho}$ using (5.7.19), one can compute all the $\omega_{\mu\nu\rho}$ using the following theorem.

Theorem 5.7.4

$$\omega_{\mu\nu\rho} = \frac{1}{2}(\Lambda_{\mu\nu\rho} + \Lambda_{\rho\mu\nu} - \Lambda_{\nu\rho\mu}). \quad (5.7.20)$$

Proof It follows from the torsion-free condition of ∇_a that the lower indices of the Christoffel symbol are symmetric, i.e., $\Gamma^\mu{}_{\nu\sigma} = \Gamma^\mu{}_{\sigma\nu}$. Hence,

$$(e_v)_{\lambda,\tau} - (e_v)_{\tau,\lambda} = (e_v)_{\lambda;\tau} - (e_v)_{\tau;\lambda},$$

and thus (5.7.19) can be rewritten as

$$\begin{aligned} \Lambda_{\mu\nu\rho} &= [\nabla_a(e_v)_b - \nabla_b(e_v)_a](\partial/\partial x^\tau)^a(\partial/\partial x^\lambda)^b(e_\mu)^\lambda(e_\rho)^\tau \\ &= [\nabla_a(e_v)_b - \nabla_b(e_v)_a](e_\rho)^a(e_\mu)^b = \omega_{\mu\nu\rho} - \omega_{\rho\nu\mu}. \end{aligned}$$

From here, it is not difficult to get (5.7.20). \square

Equation (5.7.20) is the explicit expression for $\omega_{\mu\nu\rho}$, which is convenient for calculating $\omega_{\mu\nu\rho}$ directly. However, the drawback is that this formula involves too many equations. If the metric has some symmetries, it is usually faster to find the $\omega_\mu{}^\nu$ for a rigid tetrad using Cartan's first equation of structure (see the method given after the solution of Example 1). Now we give a specific example of the calculation.

Example 1 Given the expression for the line element of a spacetime metric g_{ab} in the $\{t, r, \theta, \varphi\}$ coordinate system:

$$ds^2 = -e^{2A(r)}dt^2 + e^{2B(r)}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \quad (5.7.21)$$

find all of its curvature 2-forms $R_\mu{}^\nu$ using an orthonormal tetrad.

Solution (a) Choose an orthonormal tetrad. It follows from (5.7.21) that the coordinate basis vectors are orthogonal but not normalized; therefore, to make from them an orthonormal basis, one may choose

$$\begin{aligned} (e_0)^a &= e^{-A}(\partial/\partial t)^a, & (e_1)^a &= e^{-B}(\partial/\partial r)^a, \\ (e_2)^a &= r^{-1}(\partial/\partial\theta)^a, & (e_3)^a &= (r \sin\theta)^{-1}(\partial/\partial\varphi)^a, \end{aligned} \quad (5.7.22)$$

and the corresponding dual basis vectors are

$$\begin{aligned} (e^0)_a &= e^A(dt)_a, & (e^1)_a &= e^B(dr)_a, \\ (e^2)_a &= r(d\theta)_a, & (e^3)_a &= (r \sin\theta)(d\varphi)_a. \end{aligned} \quad (5.7.23)$$

Or, lowering the indices for (5.7.22) yields

$$\begin{aligned}(e_0)_a &= -e^A(dt)_a, & (e_1)_a &= e^B(dr)_a, \\ (e_2)_a &= r(d\theta)_a, & (e_3)_a &= (r \sin \theta)(d\varphi)_a.\end{aligned}\quad (5.7.24)$$

(b) Compute $\Lambda_{\mu\nu\rho}$ using (5.7.19). In the calculation we need a coordinate system, and naturally we choose the given system $\{t, r, \theta, \varphi\}$. Noticing the antisymmetric relation $\Lambda_{\mu\nu\rho} = -\Lambda_{\rho\nu\mu}$, one can first find all (six) independent $\Lambda_{\mu 0\rho}$ (namely, $\Lambda_{001}, \Lambda_{002}, \Lambda_{003}, \Lambda_{102}, \Lambda_{103}, \Lambda_{203}$), and then find all the independent $\Lambda_{\mu 1\rho}, \dots$. Equation (5.7.24) indicates that the only nonvanishing component of $(e_0)_\lambda$ is $(e_0)_0 = -e^A$, which is only a function of r , and hence the only nonvanishing term of $(e_0)_{0,\tau}$ is $(e_0)_{0,1} = -A'e^A$ (where ' stands for the derivative with respect to r). Thus,

$$\Lambda_{\mu 0\rho} = [(e_0)_{0,1} - 0](e_\mu)^0(e_\rho)^1 = -A'e^A(e_\mu)^0(e_\rho)^1.$$

Also, $(e_\mu)^0$ and $(e_\rho)^1$ are nonvanishing unless $\mu = 0$ and $\rho = 1$; hence, the only nonvanishing $\Lambda_{\mu 0\rho}$ is

$$\Lambda_{001} = -A'e^A(e_0)^0(e_1)^1 = -A'e^Ae^{-A}e^{-B} = -A'e^{-B}.$$

Similarly, one can find that the nonvanishing $\Lambda_{\mu\nu\rho}$ are

$$\begin{aligned}\Lambda_{001} &= -\Lambda_{100} = -A'e^{-B}, & \Lambda_{122} &= -\Lambda_{221} = -r^{-1}e^{-B}, \\ \Lambda_{133} &= -\Lambda_{331} = -r^{-1}e^{-B}, & \Lambda_{233} &= -\Lambda_{332} = -r^{-1}\cot\theta.\end{aligned}$$

Plugging into (5.7.20) yields the nonvanishing $\omega_{\mu\nu\rho}$ (note that $\omega_{\mu\nu\rho} = -\omega_{\nu\mu\rho}$):

$$\begin{aligned}\omega_{010} &= -\omega_{100} = -A'e^{-B}, & \omega_{122} &= -\omega_{212} = -r^{-1}e^{-B}, \\ \omega_{133} &= -\omega_{313} = -r^{-1}e^{-B}, & \omega_{233} &= -\omega_{323} = -r^{-1}\cot\theta.\end{aligned}$$

Therefore, the six independent connection 1-forms $\omega_{\mu\nu}$ are

$$\begin{aligned}\omega_{01} &= -A'e^{-B}e^0, & \omega_{02} &= 0, & \omega_{03} &= 0, \\ \omega_{12} &= -r^{-1}e^{-B}e^2, & \omega_{13} &= -r^{-1}e^{-B}e^3, & \omega_{23} &= -r^{-1}\cot\theta e^3.\end{aligned}$$

(c) Derive the curvature 2-forms using Cartan's second equation of structure. To find the exterior differentiation more conveniently, we rewrite the nonvanishing $\omega_{\mu\nu}$ in terms of the dual coordinate basis vectors:

$$\begin{aligned}\omega_{01} &= -A'e^{A-B}dt, & \omega_{12} &= -e^{-B}d\theta, \\ \omega_{13} &= -e^{-B}\sin\theta d\varphi, & \omega_{23} &= -\cos\theta d\varphi.\end{aligned}$$

It follows from $\omega_\mu{}^\nu = g^{\nu\sigma}\omega_{\mu\sigma} = \eta^{\nu\sigma}\omega_{\mu\sigma}$ that $\omega_0{}^i = \omega_{0i}$, $\omega_i{}^j = \omega_{ij}$ ($i, j = 1, 2, 3$). Plugging into (5.7.8), it is not difficult to find

$$\begin{aligned}\mathbf{R}_0{}^1 &= e^{-2B}(A'' - A'B' + A'^2)e^0 \wedge e^1, & \mathbf{R}_0{}^2 &= r^{-1}A'e^{-2B}e^0 \wedge e^2, \\ \mathbf{R}_0{}^3 &= r^{-1}A'e^{-2B}e^0 \wedge e^3, & \mathbf{R}_1{}^2 &= r^{-1}B'e^{-2B}e^1 \wedge e^2, \\ \mathbf{R}_1{}^3 &= r^{-1}B'e^{-2B}e^1 \wedge e^3, & \mathbf{R}_2{}^3 &= r^{-2}(1 - e^{-2B})e^2 \wedge e^3.\end{aligned}$$

Here we only give the calculation for the longest one $\mathbf{R}_1{}^3$:

$$\begin{aligned}
d\omega_1^3 &= - \left[\frac{\partial}{\partial r} (e^{-B} \sin \theta) dr + \frac{\partial}{\partial \theta} (e^{-B} \sin \theta) d\theta \right] \wedge d\varphi \\
&= e^{-B} (B' \sin \theta dr \wedge d\varphi - \cos \theta d\theta \wedge d\varphi) \\
&= r^{-1} B' e^{-2B} e^1 \wedge e^3 - r^{-2} e^{-B} \cot \theta e^2 \wedge e^3, \\
\omega_1^{\lambda} \wedge \omega_{\lambda}^3 &= \omega_1^2 \wedge \omega_2^3 = \omega_{12} \wedge \omega_{23} = r^{-2} e^{-B} \cot \theta e^2 \wedge e^3, \\
R_1^3 &= d\omega_1^3 + \omega_1^{\lambda} \wedge \omega_{\lambda}^3 = r^{-1} B' e^{-2B} e^1 \wedge e^3. \quad \blacksquare
\end{aligned}$$

We have demonstrated above the computation of the connection 1-forms ω_{μ}^{ν} using (5.7.20). Now, with the same example, we introduce an equivalent method of deriving the ω_{μ}^{ν} using Cartan's first equation of structure. Taking the exterior differentiation of (5.7.23) and plugging it into Cartan's first equation (5.7.6), we find

$$A' e^{-B} e^1 \wedge e^0 = -e^1 \wedge \omega_1^0 - e^2 \wedge \omega_2^0 - e^3 \wedge \omega_3^0, \quad (5.7.25a)$$

$$0 = -e^0 \wedge \omega_0^1 - e^2 \wedge \omega_2^1 - e^3 \wedge \omega_3^1, \quad (5.7.25b)$$

$$r^{-1} e^{-B} e^1 \wedge e^2 = -e^0 \wedge \omega_0^2 - e^1 \wedge \omega_1^2 - e^3 \wedge \omega_3^2, \quad (5.7.25c)$$

$$r^{-1} e^{-B} e^1 \wedge e^3 + r^{-1} \cot \theta e^2 \wedge e^3 = -e^0 \wedge \omega_0^3 - e^1 \wedge \omega_1^3 - e^2 \wedge \omega_2^3. \quad (5.7.25d)$$

In principle, expanding the 1-forms ω_{μ}^{ν} in terms of the basis e^{μ} (e.g., $\omega_1^0 = \alpha_0 e^0 + \alpha_1 e^1 + \alpha_2 e^2 + \alpha_3 e^3$) and plugging the result into (5.7.25), one can obtain all the ω_{μ}^{ν} . In fact, however, one can usually "read off" or even guess the correct solution in a much simpler manner. For example, the following guesses for ω_1^0 , ω_2^0 and ω_3^0 will satisfy (5.7.25a):

$$\omega_1^0 = -A' e^{-B} e^0 + \alpha_1 e^1, \quad \omega_2^0 = \omega_3^0 = 0. \quad (5.7.26)$$

Plugging the above results into (5.7.25b) yields

$$0 = -\alpha_1 e^0 \wedge e^1 - e^2 \wedge \omega_2^1 - e^3 \wedge \omega_3^1.$$

The last two terms in this equation do not contain $e^0 \wedge e^1$, and hence $\alpha_1 = 0$. It seems that one could guess $\omega_2^1 = \omega_3^1 = 0$; however, $\omega_2^1 = 0$ cannot satisfy (5.7.25c) and $\omega_3^1 = 0$ cannot satisfy (5.7.25d). From (5.7.25c) one can guess that $\omega_1^2 = -r^{-1} e^{-B} e^2$, and from (5.7.25d) one can guess that $\omega_1^3 = -r^{-1} e^{-B} e^3$ and $\omega_2^3 = -r^{-1} \cot \theta e^3$. It can be easily seen that these guesses also satisfy (5.7.25b) and (5.7.25c). Thus, the solution we just guessed, i.e.,

$$\begin{aligned}
\omega_1^0 &= -A' e^{-B} e^0, & \omega_2^0 = \omega_3^0 &= 0, \\
\omega_1^2 &= -r^{-1} e^{-B} e^2, & \omega_1^3 &= -r^{-1} e^{-B} e^3, & \omega_2^3 &= -r^{-1} \cot \theta e^3,
\end{aligned}$$

satisfies Cartan's equation, and therefore is the correct answer [which is the same as the result of step (b) in the solution of Example 1].

So far we have introduced two methods for computing the Riemann tensor R_{abc}^d : the coordinate basis method and the tetrad method (especially the orthonormal tetrad method). Each of these two methods has advantages and disadvantages, one can choose which one to use based on the specific problem and their own proficiency. Someone might wish that there is a method that combines the coordinate basis method and the orthonormal tetrad method, namely wish that there is an orthonormal coordinate basis. However, this is impossible unless g_{ab} is a flat metric. The reason is simple: the coordinate basis being orthonormal indicates that $g_{ab} = \eta_{\mu\nu} (\partial/\partial x^{\mu})^a (\partial/\partial x^{\nu})^b$. Suppose ∂_a is the ordinary derivative operator of the coordinate system, then $\partial_a g_{bc} = 0$, and hence ∂_a is the derivative operator associated with g_{ab} . Since $\partial_a \partial_b \omega_c = 0 \forall \omega_c$, we see that the R_{abc}^d for g_{ab} vanishes, i.e., g_{ab} is flat.

Exercises

- ~5.1. Complete the proof of Theorem 5.1.3 by showing that the 2-forms $(e^1)_a \wedge (e^2)_b$, $(e^2)_a \wedge (e^3)_b$ and $(e^3)_a \wedge (e^1)_b$ are linearly independent.
- ~5.2. Suppose V is a vector space and $\{(e^1)_a, (e^2)_a, (e^3)_a, (e^4)_a\}$ is a basis of V^* . Find the expansion of $\omega_a \in \Lambda(1)$, $\omega_{ab} \in \Lambda(2)$, $\omega_{abc} \in \Lambda(3)$ and $\omega_{abcd} \in \Lambda(4)$ in this basis and explain the definition of the coefficients (e.g., ω_{12}).
- ~5.3. Using mathematical induction, show that $(\omega^1)_{a_1} \wedge \cdots \wedge (\omega^l)_{a_l} = l! (\omega^1)_{[a_1} \cdots (\omega^l)_{a_l]}$, where $(\omega^1)_a, \dots, (\omega^l)_a$ are arbitrary dual vectors.
- ~5.4. Prove Theorem 5.1.4.
- ~5.5. Suppose ω is a 1-form field and u and v are vector fields. Show that $d\omega(u, v) = u(\omega(v)) - v(\omega(u)) - \omega([u, v])$. The left-hand side represents the result of $d\omega$ acting on u and v , i.e., $(d\omega)_{ab} u^a v^b$.
- ~5.6. Suppose v^b and $\omega_{a_1 \dots a_l}$ are a vector field and an l -form field, respectively, on a manifold M . Show that
- $\mathcal{L}_v \omega_{a_1 \dots a_l} = d_{a_1} (v^b \omega_{ba_2 \dots a_l}) + (d\omega)_{ba_1 \dots a_l} v^b$.
NB: Let $\mu_{a_2 \dots a_l} \equiv v^b \omega_{ba_2 \dots a_l}$, then $d_{a_1} \mu_{a_2 \dots a_l}$ means $(d\mu)_{a_1 a_2 \dots a_l}$.
 - $\mathcal{L}_v d\omega = d\mathcal{L}_v \omega$ (this is actually a very useful identity).
Hints: (1) One can first prove the special case of (a) where $l = 2$, and then it is not difficult to generalize it after getting the feeling.
(2) The result of (a) can make the proof of (b) quite simple.

- 5.7. Suppose O is the coordinate patch of the coordinate system $\{x^\mu\}$ on an n -dimensional manifold M (and O is homeomorphic to \mathbb{R}^n) and that ω_a is a 1-form field on O . Show that

$$\frac{\partial \omega_\mu}{\partial x^\nu} = \frac{\partial \omega_\nu}{\partial x^\mu} \quad (\mu, \nu = 1, \dots, n)$$

if and only if there exists $f : O \rightarrow \mathbb{R}$ such that $\nabla_a f = \omega_a$. Hint: follow the proof of Corollary 5.1.6 in Sect. 5.1.

- 5.8. Suppose $\{x, y, z\}$ and $\{r, \theta, \varphi\}$ are a Cartesian coordinate system and a spherical coordinate system, respectively, of the 3-dimensional Euclidean space. Write down the expression for $dr \wedge d\theta \wedge d\varphi$ in terms of $dx \wedge dy \wedge dz$.
- ~5.9. A connected manifold M together with a metric field g_{ab} with a Lorentzian signature is called a **spacetime**. Suppose F_{ab} is a 2-form field on an arbitrary 4-dimensional spacetime (we will see in Chap. 6 that the electromagnetic field tensor F_{ab} is exactly a 2-form field), show that

$$\frac{1}{2} (F_{ac} F_b{}^c + {}^* F_{ac} {}^* F_b{}^c) = F_{ac} F_b{}^c - \frac{1}{4} g_{ab} F_{cd} F^{cd},$$

where ${}^* F_{ac} \equiv ({}^* F)_{ac}$, ${}^* F_b{}^c = g^{ac} {}^* F_{ba}$ (this identity is helpful for studying electromagnetic fields).

- *5.10. Show that $\hat{\epsilon}_{a_1 \dots a_{n-1}} \equiv \pm n^b \hat{\epsilon}_{ba_1 \dots a_{n-1}}$ is the volume element on ∂N associated with the induced metric field h_{ab} .

- 5.11. Prove Theorems 5.6.1 and 5.6.2.
- 5.12. Suppose x, y, z are Cartesian coordinates of the 3-dimensional Euclidean space. Show that (a) ${}^*dx = dy \wedge dz$; (b) ${}^*(dx \wedge dy \wedge dz) = 1$.
- 5.13. Suppose $\{r, \theta, \varphi\}$ is a spherical coordinate system of the 3-dimensional Euclidean space. Show that, ${}^*dr = (r^2 \sin \theta)d\theta \wedge d\varphi$.
- 5.14. Suppose \vec{A} and \vec{B} are vector fields on \mathbb{R}^3 and $\vec{\nabla}$ is the derivative operator on \mathbb{R}^3 associated with the Euclidean metric. Show that
- $$\vec{\nabla} \times (\vec{A} \times \vec{B}) = (\vec{B} \cdot \vec{\nabla})\vec{A} + (\vec{\nabla} \cdot \vec{B})\vec{A} - (\vec{A} \cdot \vec{\nabla})\vec{B} - (\vec{\nabla} \cdot \vec{A})\vec{B}.$$
- 5.15. Using differential forms, prove the following well-known propositions that are not so easy to prove by the vector analysis of the 3-dimensional Euclidean space (see the end of Sect. 5.6):
- (1) a vector with no curl must be a gradient field;
 - (2) a vector with no divergence must be a curl field.
- 5.16. Suppose ∇_a is the associated derivative operator on a generalized Riemannian space (M, g_{ab}) (i.e., $\nabla_a g_{bc} = 0$), ϵ is the associated volume element (i.e., $\nabla_a \epsilon_{b_1 \dots b_n} = 0$), v^a is a vector field on M , $v_a \equiv g_{ab}v^b$ is the 1-form corresponding to v^a , and *v is the dual form field of v_a . Show that $(\nabla_a v^a)\epsilon = d^*v$. NB: This conclusion can be generalized as follows: suppose $F_{a_1 \dots a_k}$ is a k -form field ($k \leq n$), denoted by \mathbf{F} for short, and denote the $(k-1)$ -form field $\nabla^{a_k} F_{a_1 \dots a_k}$ as $\text{div } \mathbf{F}$, then ${}^*(\text{div } \mathbf{F}) = d^*\mathbf{F}$. The Maxwell equations of an electromagnetic field (see Sect. 12.6.1) provide an example.
- 5.17. Show that the $\Gamma^\sigma_{\mu\tau}$ defined by (5.7.2) are exactly the components of the Christoffel symbol defined in Sect. 3.1 with respect to the given coordinate basis in (5.7.2).
- *5.18. Using the orthonormal tetrad method, find all the tetrad components of the curvature tensors of the metrics in Exercises 14–16 of Chap. 3, and verify that the results are the same as those of the curvature tensors derived from the coordinate basis method. To distinguish from the coordinate components $R_{\mu\nu\sigma}{}^\rho$ of $R_{abc}{}^d$, one may change the notation of the tetrad components to $R_{(\mu)(\nu)(\sigma)}{}^{(\rho)}$ after obtaining all the tetrad components of $R_{abc}{}^d$.

References

- Abraham, R. and Marsden, J. (1978), *Foundations of Mechanics*, Addison-Wesley Publishing Company, Redwood City.
- Chern, S. S., Chen, W. and Lam, K. S. (1999), *Lectures on Differential Geometry*, World Scientific Publishing Company, Singapore.
- Wald, R. M. (1984), *General Relativity*, The University of Chicago Press, Chicago.
- Warner, F. W. (1983), *Foundations of Differentiable Manifolds and Lie Groups*, Springer-Verlag, New York.

Chapter 6

Special Relativity



6.1 Foundations of the 4-Dimensional Formulation

The traditional way to formulate special relativity is to use the so-called $3 + 1$ -dimensional (or, for short, 3-dimensional) formulation, in which space and time are treated separately in specific coordinate systems. However, after acquiring an understanding of differential geometry in the previous chapters, one can also use a 4-dimensional “global” way to formulate special relativity, which not only makes it easier to grasp the essence of the theory but also provides a necessary foundation for learning general relativity. The mission of this chapter is to provide this geometric reformulation of special relativity, that is, rather than using the 3-dimensional formulation, we will develop a clearer and deeper understanding by approaching the theory through the language of 4-dimensional geometry. Note that in this chapter we assume that our readers have learned the basics of special relativity.

6.1.1 Preliminaries

Physics studies the evolution of physical objects. For the convenience of study, people usually use physical models to describe physical objects. Models are the idealized version of objects, such as point masses, point charges, charged surfaces, etc.

Now let us introduce a few fundamental concepts that will later be frequently encountered using the language of models.

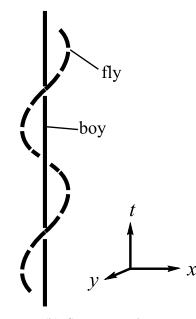
An “event” is supposed to be a very intuitive concept. A bomb explosion, a car crash, a cough are all events, each of which occupies a certain part of space and lasts for a certain period of time. The concept of an event in physics, however, is the modeling of a real event, i.e., we regard every event as happening at a point in space and an instant in time. No matter what is happening, the combination of a point in space and an instant in time is called an **event**. The collection of all the events is called a **spacetime**, and thus each event is a spacetime point. According to our

physical intuition, a spacetime should be a “4-dimensional continuum”; however, the precise definition of this phrase was not clear at first. Later, people found that the mathematical concept that can make this precise is a 4-dimensional manifold. Treating a spacetime as a 4-dimensional manifold (together with proper additional structures, e.g., a Lorentzian metric) is a basic starting point (postulate) assumed in physics. Both pre-relativity physics and special relativity assume that the spacetime manifold is \mathbb{R}^4 (the difference is their additional structures on \mathbb{R}^4 , see later); general relativity, on the other hand, allows the spacetime manifold to be any 4-dimensional connected manifold.

A point mass in Newtonian mechanics is a modeling concept that refers to a massive point in space. To discuss relativity, we generalize the concept of a point mass to a **particle**. Here a particle means a modeling particle, which is related to but different from those specific particles in physics, such as protons, neutrons, etc., in that it has no size at all. We can classify the particles into two types [see Synge (1956)]: those with (rest) mass, which are the same as point masses; and those without (rest) mass, which are often called **photons** for convenience’s sake. The whole history of a particle is formed by a series of events, and therefore corresponds to a curve in spacetime, called the **world line** of this particle. Suppose a boy (viewed as a point mass) is at rest on the ground, and a fly is experiencing a uniform circular motion around the boy (see Fig. 6.1a), then the world line of both the boy and the fly is as shown in Fig. 6.1b (called a **spacetime diagram**). In a spacetime diagram, the upward direction represents the time direction, and the horizontal directions represent spatial directions. Each horizontal slice represents the whole space at a certain moment of time. One can see the whole process of the motion (evolution) by viewing the spacetime diagram from the bottom up.

A person who makes physical measurements is called an **observer**. Usually an observer is modeled as a point mass. To make a measurement, the observer should be equipped with an accurate clock, called a **standard clock**, and the reading of this clock is called the **proper time** of this observer (see Sect. 6.1.4 for details). More generally, one can consider that any point mass carries a standard clock, and each point mass has its own proper time. Mathematically speaking, proper time is nothing but a special parameter for the world line of a point mass. An observer can only make

Fig. 6.1 A boy is at rest on the ground, around which a fly is experiencing a uniform circular motion



direct measurements of the events which happen on its own world line. In order to observe any event in the whole spacetime (or in an open subset of it), one needs to set observers everywhere (like a “patrol”), and these ubiquitous observers form a reference frame. More precisely, the set \mathcal{R} of an infinite number of observers is called a **frame of reference**, or a **reference frame**, if it satisfies the following condition: any point in spacetime (or in an open subset of spacetime) is passed through by one and only one observer in \mathcal{R} . This abstract definition is actually the specification and generalization of the often used concept of a reference frame. Take the familiar example of a moving train. Imagine the train being filled with passengers (observers), each of which carries a standard clock and is labeled by three real numbers (the spatial coordinates). Any event which happens inside the train must happen to an observer, who can record the spacetime coordinates t, x, y, z of this event (where t is the reading of the standard clock and x, y, z are the spatial coordinates of the observer). Although a train has only a limited size (length, width and height), when we talk about the “train frame”, i.e., the reference frame of the train, as a modeled concept, we have already assumed that the whole space is filled with observers. To be specific, each spatial point is occupied by an observer in the train frame; these observers move along with the train, which means they are motionless with respect to the observers inside the train. On the other hand, the observers in the “ground frame” also fill up the whole space, but they have a relative velocity with respect to the observers in the train frame. If we use vertical lines to represent the world lines of the ground frame observers in a spacetime diagram, the world lines of the train frame observers will be parallel oblique lines (the reader should draw a picture). The specification and generalization of this understanding (allowing two world lines in a frame to be non-parallel, i.e., allowing the distance between two observers to change with time) lead us to the preceding definition of a reference frame.

6.1.2 *The Background Spacetime of Special Relativity*

The so-called “geometric formulation” of special relativity actually refers to the construction of a 4-dimensional (rather than 3-dimensional) model using the language of differential geometry. The conclusions we derive will certainly agree with the 3-dimensional formulation of special relativity. To construct this geometric formulation, the first problem is: what manifold, together with what additional structure, should we use as the background spacetime? Physically speaking, any event in special relativity can be described by the coordinates of an inertial frame. The ranges for the coordinates t, x, y, z of any inertial frame are all from $-\infty$ to ∞ . Suppose p and q are two neighboring points (see Fig. 6.2), which represent two neighboring events in physics. According to special relativity, the important physical quantity that describes the relationship between p and q is the infinitesimal interval, which can be defined by means of an inertial coordinate system $\{t, x, y, z\}$ as

$$ds^2 = -dt^2 + dx^2 + dy^2 + dz^2. \quad (6.1.1)$$

[This book adopts the geometrized unit system, in which $c = 1$ (for details, see Appendix A)]. An important property of an infinitesimal interval is that it preserves its form when transformed from one inertial frame to another inertial frame, i.e.,

$$-dt^2 + dx^2 + dy^2 + dz^2 = -dt'^2 + dx'^2 + dy'^2 + dz'^2.$$

(This invariance of the interval can be verified by performing a Lorentz transformation). This reminds us of the 4-dimensional Minkowski space in mathematics: the line element in Minkowski space (\mathbb{R}^4, η_{ab}) can be expressed in terms of a Lorentzian coordinate system $\{x^0, x^1, x^2, x^3\}$ as

$$ds^2 = -(dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2. \quad (6.1.1')$$

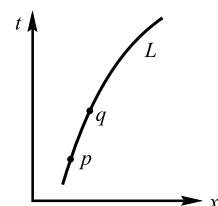
This equation has the same form as (6.1.1), and it preserves this form when transformed from one Lorentzian system to another Lorentzian system. Hence, one can see that an infinitesimal interval in physics corresponds to a Minkowski line element in mathematics, an inertial coordinate system in physics corresponds to a Lorentzian coordinate system in mathematics, and the background spacetime of special relativity corresponds to the Minkowski space (\mathbb{R}^4, η_{ab}). (Thus, a Minkowski space is also called a **Minkowski spacetime**. We may regard Minkowski space as an expression leaning towards the mathematics side, and Minkowski spacetime as leaning towards the physics side). Even further, by changing “corresponds to” to “is identical to”, one can say that the background spacetime of special relativity is Minkowski spacetime. That is, special relativity is the study regarding the evolution of physical objects in Minkowski spacetime. Any physical phenomenon happening in Minkowski spacetime belongs to the scope of special relativity.

Using an inertial coordinate system, one can define the speed of any particle. Suppose L is the world line of a particle, p and q are two neighboring points on L (see Fig. 6.2), and (t_1, x_1, y_1, z_1) and (t_2, x_2, y_2, z_2) are the coordinates of p and q in an inertial frame \mathcal{R} . Let

$$dt \equiv t_2 - t_1, \quad dx \equiv x_2 - x_1, \quad dy \equiv y_2 - y_1, \quad dz \equiv z_2 - z_1,$$

then the speed of the particle at p relative to the frame \mathcal{R} is defined as

Fig. 6.2 The world line of a particle and a line segment



$$u := \frac{\sqrt{dx^2 + dy^2 + dz^2}}{dt}. \quad (6.1.2)$$

Hence, it follows from (6.1.1) that the line element of L in between p and q is

$$ds^2 = -(1 - u^2)dt^2. \quad (6.1.3)$$

The equation above indicates that $u = 1$ is equivalent to $ds^2 = 0$ (line element being null); $u < 1$ is equivalent to $ds^2 < 0$ (line element being timelike). Therefore, the two significant basic tenets of special relativity expressed in the $3 + 1$ -dimensional formulation—① the speed of a photon relative to any inertial frame is $u = 1$; ② the speed of a point mass relative to any inertial frame is $u < 1$ —can now be reformulated in terms of the 4-dimensional language as follows:

- ① **The world line of a photon is a null curve in Minkowski spacetime;**
- ② **The world line of a point mass is a timelike curve in Minkowski spacetime.**

In the $3 + 1$ -formulation one always needs a reference frame, and these basic tenets above also require a definition of the speed u . Both “a reference frame” and “a definition of speed” depend on one’s own choice, and therefore belong to “human factors”. The use of these “human factors” not only makes it fail to be as concise and self-contained as the 4-dimensional formulation, but sometimes may also lead to misunderstanding. For instance, if we make different definitions of the speed (there are several that, in some sense, are qualified to be called as “speed”), then a “faster-than-light” particle would not contradict with the basic tenets in bold above (the world line of a “faster-than-light” point mass can still be a timelike curve). For instance, an important example of “faster-than-light” speed that does not violate relativity is the recessional velocity of a galaxy, which will be discussed in detail in Sect. 10.2.1. However, if one had only heard that “relativity does not allow travel at a speed faster than light,” then they might think naively that this kind of seemingly “faster-than-light” travel is forbidden by relativity. Nevertheless, a point mass whose world line is a spacelike curve is certainly forbidden by relativity. Therefore, we can see that the 4-dimensional geometric formulation naturally clarifies what is and is not a violation of special relativity.

6.1.3 Inertial Observers and Inertial Frames

The fundamental postulates of special relativity are: the principle of invariant light speed; and the special principle of relativity. The latter further contains the following two aspects.

- ① Among all observers (i.e., point masses), there exists a special kind of observer, called **inertial observers**, which are essentially distinguished from all the other observers (non-inertial observers); that is, one can choose a special subset from the collection of all the observers, in which each element is an inertial observer.

② All inertial observers are on an equal footing, i.e., no inertial observer is preferred over any other; that is, one cannot choose a special element (or several) from the subset formed by inertial observers. For example, one cannot ask which inertial observer is at absolute rest.

Now we discuss the mathematical counterpart for an inertial observer. According to the 3-dimensional formulation of special relativity, the speed of an inertial observer relative to its own inertial coordinate system $\{t, x, y, z\}$ is $u = 0$, and thus its world line coincides with a t -coordinate line in this system. Suppose ∂_a is the ordinary derivative operator of this system, then $\partial_a(\partial/\partial t)^b = 0$. Hence,

$$\left(\frac{\partial}{\partial t}\right)^a \partial_a \left(\frac{\partial}{\partial t}\right)^b = 0. \quad (6.1.4)$$

Noting that an inertial coordinate system is a Lorentzian coordinate system, we see that ∂_a is the derivative operator associated with the Minkowski metric η_{ab} (satisfying $\partial_a \eta_{bc} = 0$). Thus, (6.1.4) is a geodesic equation of Minkowski space, and hence the world line of any inertial observer is a timelike geodesic. On the other hand, it can also be proved that for any given timelike geodesic G one can always find a Lorentzian coordinate system such that G is a t -coordinate line, and thus G represents an inertial observer. Therefore, an inertial observer in physics corresponds to a timelike geodesic in mathematics, or one can say that the world line of an inertial observer is a timelike geodesic. From the mathematical perspective, timelike geodesics are the most natural and simplest type of timelike curves; from the physical perspective, inertial observers are the most natural and simplest type of observers. This is an elegant correspondence between inertial observers and timelike geodesics.

Since each t -coordinate line in a Lorentzian coordinate system corresponds to an inertial observer, the reference frames formed by all the t -coordinate lines in this system is called an **inertial reference frame**, and this coordinate system is called an **inertial coordinate system** in this inertial reference frame. When it is not necessary to distinguish a reference frame and a coordinate system, both an inertial reference frame and an inertial coordinate systems are called an **inertial frame** for short. The domain of an inertial frame is the whole spacetime (the whole \mathbb{R}^4), and thus is also called a global inertial frame. The world lines of all the observers in the same inertial frame are parallel geodesics; in contrast, if two inertial observes belong to two different inertial frame (such as the “train frame” and the “ground frame” we mentioned above), then their world lines are geodesics that are not parallel to each other. A point mass is said to be “free” if its world line is a geodesic, i.e., it is undergoing inertial motion.

According to Theorem 4.3.6, a coordinate transformation between two Lorentzian systems in the 4-dimensional Minkowski spacetime $(\mathbb{R}^4, \eta_{ab})$ corresponds to an isometry in $(\mathbb{R}^4, \eta_{ab})$. Any isometry can be constructed from several basic isometries, the latter of which includes the isometries that are “continuous” and “discrete”. The “discrete” ones are reflections and inversions, while the “continuous” ones includes three types [see Sect. 4.3, Example 1(4)]: (a) translations, represented by

4 independent Killing vector fields: $(\partial/\partial t)^a$, $(\partial/\partial x)^a$, $(\partial/\partial y)^a$, $(\partial/\partial z)^a$; (b) spatial rotations, represented by 3 independent Killing vector fields: $-y(\partial/\partial x)^a + x(\partial/\partial y)^a$, $-z(\partial/\partial y)^a + y(\partial/\partial z)^a$, $-x(\partial/\partial z)^a + z(\partial/\partial x)^a$; and (c) boosts, represented by 3 independent Killing vector fields: $t(\partial/\partial x)^a + x(\partial/\partial t)^a$, $t(\partial/\partial y)^a + y(\partial/\partial t)^a$, $t(\partial/\partial z)^a + z(\partial/\partial t)^a$. Now we interpret the physical meaning of these three types of transformations by providing an example for each of them.

(a) Without loss of generality, consider time translation. In this case, the coordinate transformation induced by the one-parameter group of isometries corresponding to the Killing field $(\partial/\partial t)^a$ is

$$t' = t + a, \quad x' = x, \quad y' = y, \quad z' = z,$$

where a serves as the parameter for this one-parameter group. Physically, this transformation corresponds to adding a value a to the initial setting of the standard clocks of all the observers in the inertial frame \mathcal{R} . Daylight saving time is an example of it, where $a = 1$ (hour).

(b) Consider a rotation in the xy -surface. The coordinate transformation induced by the one-parameter group of isometries corresponding to the Killing field $-y(\partial/\partial x)^a + x(\partial/\partial y)^a$ is

$$t' = t, \quad x' = x \cos \alpha - y \sin \alpha, \quad y' = x \sin \alpha + y \cos \alpha, \quad z' = z,$$

where α is a constant that serves as the parameter. Physically, this corresponds to a spatial coordinate rotation inside the inertial reference frame.

(c) Consider a boost in the tx -surface. The coordinate transformation induced by the one-parameter group of isometries corresponding to the Killing field $t(\partial/\partial x)^a + x(\partial/\partial t)^a$ is (see Theorem 4.3.5)

$$t' = \gamma(t - vx), \quad x' = \gamma(x - vt), \quad y' = y, \quad z' = z, \quad (6.1.5)$$

where v is a constant that serves as the parameter, and $\gamma \equiv (1 - v^2)^{-1/2}$. Physically, this corresponds to the Lorentzian transformation between two reference frames \mathcal{R} and \mathcal{R}' . The coordinate axes of these two reference frames are parallel and oriented so that the frame \mathcal{R}' is moving in the positive (or negative) x -direction with a constant speed $|v|$, and the origins of their spatial coordinates are coincident at $t = t' = 0$.

Both translations and spatial rotations correspond to coordinate transformations in the same inertial reference frame. For example, a time translation is just a resetting of the zero for each standard clock owned by each observer in the same inertial reference frame; neither the observers nor the reference frame are changed. For another example, after a spatial rotation for $\{t, x, y, z\}$, the new coordinate system $\{t' = t, x', y', z'\}$ will still be an inertial coordinate system in this reference frame. Thus, there exist many inertial coordinate systems in the same inertial reference frame. Two inertial coordinate systems related by a boost, however, must belong to two different inertial reference frames since their t -coordinate lines are different.

6.1.4 Proper Time and Coordinate Time

The proper time of an observer (a point mass) is the reading of his standard clock. However, what exactly is a standard clock? We will need to add the following definition:

Definition 1 A clock is called a **standard clock** or **ideal clock** if the difference between the two readings τ_1 and τ_2 at two arbitrary points p_1 and p_2 on its world line equals the arc length of its world line between p_1 and p_2 , i.e.,

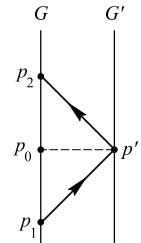
$$\tau_1 - \tau_2 = \int_{p_1}^{p_2} \sqrt{-ds^2}. \quad (6.1.6)$$

Remark 1 If we do not take $c = 1$, then the right-hand side of the above equation should be multiplied by $1/c$.

Remark 2 One should distinguish two concepts related to clocks—**rate** and (initial) **setting**. A standard clock only has a requirement on its rate (i.e., the difference of the readings at any two points on the world line equals the arc length), while the synchronization problem in a reference frame only involves the initial (zero) setting. Many regions in the world use daylight saving time, which stipulate the clock to be “one hour faster” at a certain date of each year. The word “faster” may be misunderstood as raising the rate, but it actually just means changing the setting.

Remark 3 According to Definition 1, the proper time of an observer is equal to the arc length of its world line. The zero of τ on the world line only depends on the setting, which is arbitrary when there is only one observer (or a few observers). However, if we consider a reference frame, then the zero of the proper time of each observer needs to satisfy a certain kind of requirement. For instance, suppose \mathcal{R} is an inertial reference frame, and G is one of its observers. Let any $p_0 \in G$ be the zero of the proper time of G and Σ_0 represent the hypersurface passing through p_0 that is orthogonal to the world lines of all the observers, then any observer G' in \mathcal{R} must choose the intersection of Σ_0 and their world line as the zero of their proper time. This kind of requirement is called **clock synchronization** in an inertial frame. At first sight, it seems that this could be realized as follows: Alice (observer G) sets her clock to zero, denoted by event p_0 , and simultaneously tells Bob (observer G') “set your clock to zero right now.” However, since it takes time for a signal to propagate, if we use q to represent the event of Bob receiving the notice, then q cannot be on the hypersurface Σ_0 . If Bob follows the order, i.e., sets his clock to zero at event q , then it certainly cannot satisfy the requirement of clock synchronization. Thus, we can see that clock synchronization is a nontrivial process in relativity. Here we introduce a method of synchronization. First, Alice should tell Bob beforehand, “take a mirror with you and zero your clock when you see the light signal I send.” At a point (event) p_1 , Alice would send a light signal to Bob; the light will be reflected when it arrives

Fig. 6.3 Method of clock synchronization



at Bob's mirror (event p') and Alice will see this reflected light when she is at point p_2 (see Fig. 6.3). To synchronize her clock with that of Bob, Alice just needs to zero her clock at p_0 , namely the midpoint of $p_1 p_2$ (measured by the arc length). Note that in this method we have used the fact that the speed of light does not depend on the direction (the path of a photon is a null geodesic).

Remark 4 A standard clock is also a model. What kind of real clock can be regarded as a standard clock? Experiments show that, in most cases, atomic clocks can be treated as standard clocks to a high degree of accuracy, and even the clocks in our daily life provide a good approximation. However, any real clock will deviate substantially from a standard clock in certain special cases [see Misner et al. (1973) pp. 393–395; Rindler (1982) p. 31]. For example, a pendulum clock, the mechanism of which highly depends on the gravitational acceleration of Earth, will become completely useless in a spaceship far away from the Earth. Nonetheless, this only affects the choice of a clock in an experiment and is thus irrelevant to our theoretical discussions. In theory, all we need is the concept of a standard clock.

Remark 5 Later, when we talk about a world line, we always assume that we are using the proper time τ as the parameter. Since the proper time is equal to the arc length of the world line, the length of its tangent vector $(\partial/\partial\tau)^a$ is 1 (see the paragraph before Definition 7 in Sect. 2.5). Thus, one should interpret an observer as a timelike curve with a unit tangent vector field.

Remark 6 Photons do not have the notion of proper time (the length of a null curve vanishes), and therefore cannot serve as observers.

Suppose x^0 is the timelike coordinate of a coordinate system [i.e., $\eta_{ab}(\partial/\partial x^0)^a (\partial/\partial x^0)^b < 0$], and x^1, x^2, x^3 are spacelike coordinates [i.e., $\eta_{ab}(\partial/\partial x^i)^a (\partial/\partial x^i)^b > 0, i = 1, 2, 3$], then the value of x^0 for any point p in the coordinate patch is called the **coordinate time** of an event p in this system. The coordinate time for an inertial reference frame is called an **inertial coordinate time**, whose domain is the whole \mathbb{R}^4 . One should pay close attention to the following two differences between the coordinate time and the proper time:

① Proper time only makes sense in relation to the points on the world line, and so without a world line one cannot talk about proper time. If two world lines L_1 and L_2 intersect at p , then p 's proper time on L_1 can be different from its proper time on

L_2 . In contrast, coordinate time does not depend on a world line. As long as p is a point in the coordinate patch, we can talk unambiguously about its coordinate time in this system.

② The same spacetime point p can have different coordinate times in different coordinate systems, while the proper time of an observer at p is independent of the coordinate system.

The following proposition provides the relation between the proper time and inertial coordinate time on a timelike curve.

Proposition 6.1.1 *Suppose $L(\tau)$ is the world line of a point mass, τ is the proper time, and t is the coordinate time of an inertial frame \mathcal{R} then*

$$\frac{dt}{d\tau} = \gamma_u, \quad (6.1.7)$$

where $\gamma_u \equiv (1 - u^2)^{-1/2}$, with u being the speed of the point mass relative to \mathcal{R} .

Proof Again, we use Fig. 6.2. It follows from $d\tau = \sqrt{-ds^2}$ and (6.1.3) that $d\tau^2 = (1 - u^2)dt^2$, and hence we have (6.1.7). \square

If $L(\tau)$ is a t -coordinate line in the inertial frame \mathcal{R} , then from (6.1.2) we see that $u = 0$, and hence it follows from (6.1.7) that $dt = d\tau$. Thus, the coordinate time for an inertial observer in their inertial frame is equal to their proper time.

6.1.5 Spacetime Diagrams

Diagrams are used frequently in the study of motion. The diagrams people usually use to show spatial trajectories are spatial diagrams. For example, the spatial trajectory for a projectile is a parabola. This kind of diagram does not have time involved, and cannot reflect at which point the object is located at a certain moment of time. A spacetime diagram, however, can overcome this drawback. It uses points to represent events, and curves to represents the motion (evolution) of a particle in spacetime, etc. If we only consider 1-dimensional motion, then we only need to draw a 2-dimensional spacetime diagram. When drawing the diagram, one should choose an arbitrary inertial frame \mathcal{R} , and then draw a vertical axis pointing upward as its t -axis (this axis represents the flow of time), and a horizontal axis as its x -axis (see Fig. 6.4). All kinds of particles moving along the x -axis can be represented by the curves in the figure. For example, the t -axis represents the world line of the observer G_0 at $x = 0$ in the frame \mathcal{R} , another vertical line in the picture represents the observer G_1 at $x = x_1$ in the frame \mathcal{R} (vertical indicates that the observer is at rest relative to frame \mathcal{R}), while the dashed line in the figure represents the world line of a photon. For any given moment \hat{t} , we have a point (\hat{t}, \hat{x}) on the line, whose spatial coordinate \hat{x} reflects the position of the photon at \hat{t} . What does the tilted line G'_0 in Fig. 6.4 stand for? Since it is tilted, its x -coordinate will change linearly with the t -coordinate. It

Fig. 6.4 A 2-dimensional spacetime diagram based on the reference frame \mathcal{R}

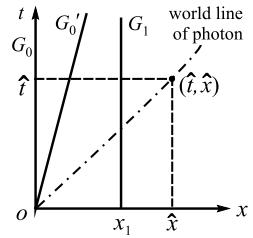
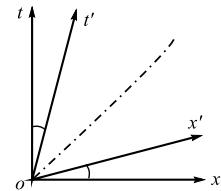


Fig. 6.5 x' - and t' -axes lay on two different sides of the 45° line



follows from (6.1.2) that its speed relative to \mathcal{R} is a constant less than 1, and thus it is a point mass experiencing inertial motion.

Actually, based on the fact that G'_0 is a timelike straight line (geodesic), we can directly tell that G'_0 is an inertial observer. Also, from the fact that G'_0 passes through the origin we can see that it is the observer at $x' = 0$ in the frame \mathcal{R}' obtained by the transformation (6.1.5), i.e., G'_0 is the t' -coordinate of \mathcal{R}' . This conclusion can also be verified another way: plugging $x' = 0$ into the Lorentz transformation (6.1.5) yields $t = x/v$, and thus the t' -axis is a straight line passing through the origin, with a slope $1/v$. How do we draw the x' -axis of the frame \mathcal{R}' ? The x' -axis satisfies $t' = 0$, and plugging this into (6.1.5) yields $t = vx$, and thus the x' -axis is a straight line passing through the origin that has a slope v ; the dashed line bisects the angle between the x' -axis and the t' -axis (see Fig. 6.5). One may ask: does this indicate that the x' -axis and t' -axis are not orthogonal, and therefore \mathcal{R} is preferred over \mathcal{R}' ? This actually is the “deception of the spacetime diagram” that comes from our Euclidean way of thinking. In fact, noticing that $\{t', x', y', z'\}$ is also a Lorentzian system, naturally we have $\eta_{ab}(\partial/\partial t')^a(\partial/\partial x')^b = 0$, i.e., $(\partial/\partial t')^a$ and $(\partial/\partial x')^b$ are orthogonal measured by the Minkowski metric. So \mathcal{R} and \mathcal{R}' are still on an equal footing. Indeed, when drawing a picture, we usually choose a reference frame first, and set their t -axis and x -axis to be, respectively, vertical and horizontal; however, the choice of this reference frame is totally arbitrary. For instance, if we choose \mathcal{R}' first, then the spacetime diagram will look like Fig. 6.6, which seems to be different from Fig. 6.5, but essentially they are the same.

The “deception” of a spacetime diagram is not only manifested in the orthogonality, but also in the judgement of length. Suppose $p = (t, x)$ is an arbitrary spacetime point, op is the straight line segment between o and p (see Fig. 6.7), whose length measured by the Minkowski metric is $l_{op} = \sqrt{|-t^2 + x^2|}$. Thus, the straight line segment between o and each point on the hyperbola $-t^2 + x^2 = K$ (constant) has the same length, e.g., $l_{op} = l_{oq}$, even though intuitively (i.e., according to the Euclidean

Fig. 6.6 The spacetime diagram based on frame \mathcal{R}' , equivalent to Fig. 6.5

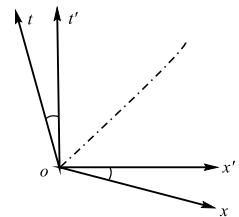


Fig. 6.7 The length of the line between o and a point on the hyperbola is a constant

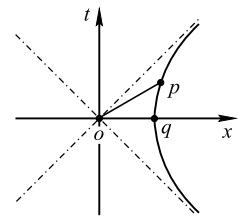
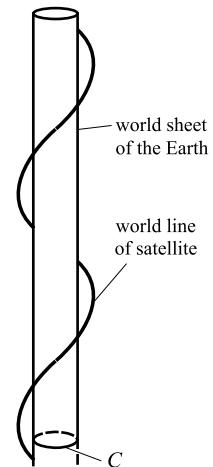


Fig. 6.8 The Earth's world sheet (with one dimension suppressed) and the satellite's world line

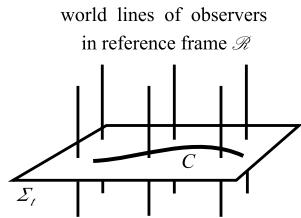


metric) their lengths are not the same in the diagram. The hyperbola in Fig. 6.7 is called a **calibration curve**.

If the physical phenomenon also involves the second and the third spatial dimensions, then the 2-dimensional spacetime diagram will not be enough. However, even if we draw in perspective we can only represent three dimensions on a piece of paper, and since we need one dimension to represent time, there are only two dimensions left; therefore, one spatial dimension cannot be represented on paper (and has to be “suppressed” in the diagram).

Luckily, in many cases, there will be one dimension (or even two) that are not important, or there exists some kind of symmetry that allows us to suppress one dimension without losing anything useful. Take an artificial satellite rotating around the Earth as an example (see Fig. 6.8). The surface of the Earth is a 2-dimensional

Fig. 6.9 A surface of simultaneity of an inertial frame \mathcal{R}



sphere; however, one dimension is suppressed when we draw the diagram, so at each moment the surface of the Earth is represented by a circle (C in the figure). Approximately, the Earth can be considered as undergoing inertial motion, which we can exploit to draw a diagram. The world line of each point on the ground (such as Beijing or New York) is a vertical line, all of which together form a cylinder, called the **world sheet** of the Earth's surface. The helix in the figure represents the world line of the satellite, the slope of which reflects the speed of its rotation.

Suppose \mathcal{R} is an inertial reference frame in Minkowski spacetime, then each point on a 3-dimensional plane (hyperplane) Σ_t , that is orthogonal to all the observers has the same t coordinate, and thus Σ_t is called a surface of simultaneity of the frame \mathcal{R} (see Fig. 6.9), which represents the “whole space” for \mathcal{R} at t . Suppose C is a curve on Σ_t , then any line segment will have $dt = 0$, and hence the Minkowski line element $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2$ will induce a spatial line element $d\hat{s}^2 = dx^2 + dy^2 + dz^2$, namely the Euclidean line element. Thus, the space at anytime for an inertial frame \mathcal{R} is a 3-dimensional Euclidean space; this is exactly what is assumed in the 3-dimensional formulation of special relativity. If we change to another inertial frame \mathcal{R}' , since the world lines of its observers are not perpendicular to the world lines of the observers in \mathcal{R} , the surfaces of simultaneity of \mathcal{R}' are certainly different from the surfaces of simultaneity of \mathcal{R} . This can be regarded as the cause of the relativity of simultaneity.

6.1.6 Spacetime Structure: Special Relativity Versus Pre-Relativity Physics

In pre-relativity physics, space and time are the most primary concepts that everyone knows. From the historical perspective, the concept of space and time came first, and then, after the birth of relativity, the concept of spacetime was gradually developed. Many people would think spacetime is not difficult to understand since “it is nothing but space and time”. However, in relativity spacetime itself is the most primary concept, while space and time are relative notions derived from spacetime. By “derived” we mean the notions of space and time only come from applying a “3 + 1” decomposition to spacetime using a reference frame, and by “relative” we mean there exist many different ways of 3 + 1 decomposition for a spacetime (Fig. 6.5 represents two

Fig. 6.10 Surfaces of absolute simultaneity in pre-relativity physics



different ways of decomposition for Minkowski spacetime using the reference frames \mathcal{R} and \mathcal{R}'). From the viewpoint of 4-dimensional geometry, the difference between the concepts of space and time in relativity and pre-relativity physics come from the difference between their spacetime structures. Pre-relativity physics assumes that the spacetime manifold is \mathbb{R}^4 , equipped with some intrinsic additional structures. The first one is a smooth function $t : \mathbb{R}^4 \rightarrow \mathbb{R}$, called the **absolute time**, such that \mathbb{R}^4 can be foliated into infinitely many slices. Each slice is a constant- t surface Σ_t (a hypersurface in \mathbb{R}^4 , see Fig. 6.10), called a **surface of absolute simultaneity**, with a 3-dimensional Euclidean metric, which represents the “whole 3-dimensional space” at t (see Optional Reading 6.1.6 for details). All the points on the same Σ_t represent the events happening simultaneously at different places, and points on different Σ_t represent the events happening at different times. The so-called “absolute simultaneity” means that simultaneity holds in whatever reference frame, which is obviously different from relativity. In special relativity, two simultaneous events in one reference frame can be non-simultaneous in another reference frame. There are only surfaces of relative simultaneity in special relativity. If we compare surfaces of simultaneity to playing cards, then pre-relativity physics only contains one deck of cards (which is independent of the reference frame, and thus is absolute) while special relativity has infinitely many decks of cards (which depend on reference frames, and thus are relative; each individual card represents the whole space at a given time in a given reference frame). This is a significant difference between the spacetime structures of these two kinds of theory.

Now we discuss the difference between these two kinds of spacetime structure from the perspective of causality. Given an event $p \in \mathbb{R}^4$, one can always write $\mathbb{R}^4 - \{p\}$ as the union of three nonintersecting subsets M_1 , M_2 and M_3 , i.e., $\mathbb{R}^4 - \{p\} = M_1 \cup M_2 \cup M_3$, where

$$\begin{aligned} M_1 &\equiv \{q \in \mathbb{R}^4 - \{p\} \mid \text{there exists an observer that experiences } q \text{ before } p\}, \\ M_2 &\equiv \{q \in \mathbb{R}^4 - \{p\} \mid \text{there exists an observer that experiences } p \text{ before } q\}, \\ M_3 &\equiv \{q \in \mathbb{R}^4 - \{p\} \mid \text{there is no observer that experiences both } p \text{ and } q\}. \end{aligned}$$

Pre-relativity physics assumes that the subset M_3 is the surface Σ_t of absolute simultaneity (with p being removed) passing through p , while M_1 and M_2 are respectively the “upper half of \mathbb{R}^4 ” and the “lower half of \mathbb{R}^4 ” on different sides of Σ_t (see Fig. 6.11). Their physical meanings are: if $q \in M_2$, then we say that the event q happens in the future of p ; if $q \in M_1$, then we say that q happens in the past of p . However, in special relativity, since the world lines of observers are timelike curves, M_2 and M_1 are respectively the subset enclosed by the future light cone surface (a null hypersurface) of p and the past light cone surface of p (excluding the points

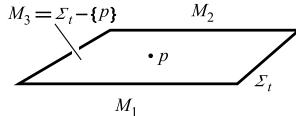


Fig. 6.11 The spacetime structure of pre-relativity physics. The surface of absolute simultaneity passing through p is a 3-dimensional surface, above and below which are the future and the past of p

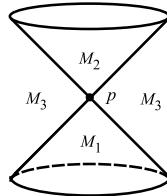


Fig. 6.12 The spacetime structure of special relativity. There is no surface of absolute simultaneity. The future and past of p are much smaller than the corresponding subset in Fig. 6.11, while the subset M_3 that has no causal relation with p is much larger than the M_3 in Fig. 6.11

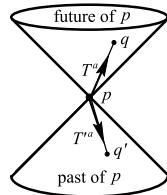


Fig. 6.13 The timelike vector T^a is future-directed, while T'^a is past-directed

on the surface), while M_3 will be a lot “bigger” than the 3-dimensional submanifold Σ_t in Fig. 6.11, which contains all the points that are not contained by M_1 and M_2 (including the points on the light cone surfaces), see Fig. 6.12.

Suppose $q \in M_2$, then the tangent vector T^a of the geodesic from p to q at p must be timelike. Similarly, if $q' \in M_1$, then the tangent vector T'^a of the geodesic from p to q' at p must also be timelike. However, physically T^a and T'^a are quite different after all: T^a is future-directed while T'^a is past-directed (see Fig. 6.13). In relativity, T^a and T'^a are called a **future-directed timelike vector** and a **past-directed timelike vector**, respectively. A timelike vector at p is either future-directed or past-directed. The (nonvanishing) future-directed and past-directed null vectors can be defined similarly.

[Optional Reading 6.1.1]

It is instructive to compare the spacetime structures of special relativity and general relativity with that of pre-relativity physics. According to general relativity, gravity in essence is the wrapping of 4-dimensional spacetime (see Sect. 7.1). Special relativity is about the physics when gravity is not present (or can be ignored), and thus the background spacetime is $(\mathbb{R}^4, \eta_{ab})$. General relativity is about the physics when there is gravity, the background spacetime of which is an arbitrary (connected) 4-dimensional manifold M together with a

curved metric field g_{ab} , i.e., (M, g_{ab}) . The background spacetime of pre-relativity physics can be revisited by formulating Newton's theory of gravity using the 4-dimensional geometric language. Based on Newton's theory of gravity, the gravitational field of the space can be described by the gravitational potential ϕ , whose relation with the mass density μ satisfies the Poisson equation

$$\nabla^2 \phi = 4\pi\mu. \quad (6.1.8)$$

[We have set the gravitational constant $G = 1$, as we adopt the geometrized unit system]. A point mass that is not subjected to any forces other than gravity is called a **free point mass**. A free point mass with unit mass obeys the following equation of motion:

$$\frac{d^2x^i}{dt^2} = -\frac{\partial\phi}{\partial x^i}, \quad i = 1, 2, 3, \quad (6.1.9)$$

where t is the Newtonian absolute time, and x^i are the spatial Galilean coordinates (i.e., the Cartesian coordinates in mathematics). After the initial conditions are given, the solutions for $x^i(t)$ in (6.1.9) can be viewed as the parametric representation of a curve in space with t as the parameter, which represents the spatial trajectory of the mass point. For instance, the trajectory of a point mass projectile near the ground is a parabola. Cartan et al. reformulated the facts above using the geometric language, the key points are as follows [also see Misner et al. (1973), Chap. 12]:

The background spacetime of Newton's theory of gravity is called **Newtonian spacetime**, which is formed by a manifold \mathbb{R}^4 and the following additional structures: (a) there exists a smooth function $t : \mathbb{R}^4 \rightarrow \mathbb{R}^4$, called the **absolute time**, satisfying certain conditions; (b) there exists a derivative operator ∇_a on \mathbb{R}^4 , whose Christoffel symbols in a given coordinate system $\{x^\mu\}$ satisfy

$$\Gamma^i_{00} = \frac{\partial f}{\partial x^i}, \quad i = 1, 2, 3 \quad (f \text{ is a function on } \mathbb{R}^4), \quad \text{the other } \Gamma^\mu_{\nu\sigma} = 0. \quad (6.1.10)$$

From these two points, one finds the following:

(1) The existence of the absolute time provides an absolute “stratification” to the spacetime manifold \mathbb{R}^4 : $\forall p \in \mathbb{R}^4$, there exists a constant- t surface Σ_t (a hypersurface in \mathbb{R}^4) such that $p \in \Sigma_t$ (see Fig. 6.10), which represents the “whole 3-dimensional space” at t , called a **surface of absolute simultaneity**. Events p and q are said to be simultaneous if $t(p) = t(q)$.

Suppose $\gamma(\lambda)$ is an arbitrary geodesic in Newtonian spacetime (λ is an affine parameter), then its parametric representation $x^\mu(\lambda)$ under a coordinate system satisfying (6.1.10) obeys the following equations:

$$\frac{d^2x^\mu}{d\lambda^2} + \Gamma^\mu_{\nu\sigma} \frac{dx^\nu}{d\lambda} \frac{dx^\sigma}{d\lambda} = 0, \quad \mu = 0, 1, 2, 3. \quad (6.1.11)$$

Let $\mu = 0$, it follows from $\Gamma^0_{\nu\sigma} = 0$ ($\nu, \sigma = 0, 1, 2, 3$) that $0 = d^2x^0/d\lambda^2 = d^2t/d\lambda^2$, and hence

$$t = \alpha\lambda + \beta, \quad \alpha, \beta \text{ are constants.} \quad (6.1.12)$$

This equation indicates that the absolute time t can serve as the affine parameter of any geodesic whose $\alpha \neq 0$. Now let $\mu = i$ in (6.1.11), it follows from (6.1.12) and $\Gamma^i_{00} = \partial f / \partial x^i$, $\Gamma^i_{jk} = 0$ ($i, j, k = 1, 2, 3$) that

$$\frac{d^2x^i}{dt^2} + \frac{\partial f}{\partial x^i} = 0, \quad i = 1, 2, 3. \quad (6.1.9')$$

Comparing this with (6.1.9) we know that, as long as we interpret f as the gravitational potential ϕ and interpret x^i as the Galilean coordinates, then a geodesic with the absolute

time t as its affine parameter in Newtonian spacetime corresponds to the world line of a free point mass.

(3) Plugging (6.1.10) into (3.4.20') and (3.4.21), it is not difficult to find the components of the Riemann tensor and Ricci tensor of ∇_a as follows (f has been changed to ϕ):

$$R_{0i0}{}^j = -R_{i00}{}^j = \frac{\partial^2 \phi}{\partial x^i \partial x^j}, \quad \text{all the other } R_{\mu\nu\rho}{}^\sigma = 0, \quad (6.1.13)$$

$$R_{00} = \sum_{i=1}^3 \frac{\partial}{\partial x^i} \frac{\partial \phi}{\partial x^i} = \nabla^2 \phi = 4\pi \mu, \quad \text{all the other } R_{\mu\nu} = 0. \quad (6.1.14)$$

Equation (6.1.13) indicates that Newtonian spacetime is not flat. (In comparison, in Einstein's theory a spacetime with gravity is not flat either). However, the derivative operator $\hat{\nabla}_a$ induced by ∇ on each surface Σ_t of simultaneity is flat. [(6.1.10) indicates that $\Gamma^i{}_{jk} = 0$, $i, j, k = 1, 2, 3$, and the corresponding 3-dimensional Riemann tensor vanishes]. This can also be verified from another point of view: when the α in (6.1.12) vanishes, the geodesic $\gamma(\lambda)$ lies on the surface Σ_β of simultaneity at $t = \beta$; from (6.1.11), $\Gamma^i{}_{jk} = 0$ and with $t = \beta$ we get $d^2x^i/d\lambda^2 = 0$, and thus

$$x^i(\lambda) = \alpha^i \lambda + \beta^i, \quad \alpha^i, \beta^i \text{ are constants.} \quad (6.1.15)$$

This is a set of linear equations, and, as long as we interpret x^i as the Cartesian coordinates of Σ_β , then (6.1.15) indicates that the geodesic on Σ_β is a straight line. In fact, one just needs to define the Euclidean metric on Σ_β in terms of x^i as follows:

$$\delta_{ab} = \delta_{ij} (dx^i)_a (dx^j)_b,$$

then the ordinary derivative operator ∂_a of the system $\{x^i\}$ will naturally satisfy $\partial_a \delta_{bc} = 0$, and the Christoffel symbols of ∂_a in $\{x^i\}$ will certainly vanish, i.e.,

$$\Gamma^i{}_{jk} = 0, \quad i, j, k = 1, 2, 3.$$

Thus, ∂_a is exactly the $\hat{\nabla}_a$ on Σ_β induced by ∇_a . Therefore, each surface Σ_t of absolute simultaneity is a 3-dimensional Euclidean space, and the Galilean coordinates used commonly in physics are exactly the Cartesian coordinates of this space. $\{t, x^i\}$ is also called a 4-dimensional Galilean coordinate system.

Here is a natural question to ask: can we define a metric on \mathbb{R}^4 that is associated with ∇_a ? The answer is negative: as long as gravity exists, the “metric” one finds must be degenerate [the signature for the “metric” with upper indices is $(0, +, +, +)$]. This optional reading provides an example for the physical application of a manifold without metric but with a derivative operator (and thus with curvature defined).

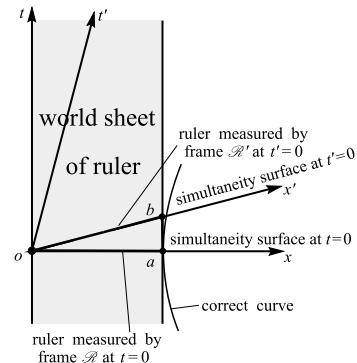
[The End of Optional Reading 6.1.1]

6.2 Interesting Typical Effects

6.2.1 Length Contraction

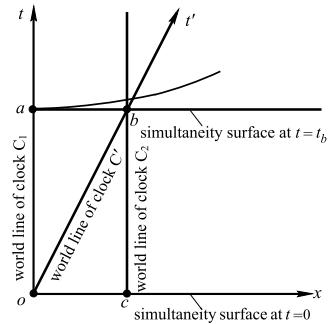
In the 3-dimensional language, a point mass is a point in space, but in the 4-dimensional language it is a timelike curve (world line) in spacetime. Similarly, a ruler in the 3-dimensional language is a line segment in space, while in the

Fig. 6.14 The world sheet of the ruler is absolute; the line segments oa and ob are the 1-dimensional rulers observed in \mathcal{R} and \mathcal{R}' at $t = 0$ and $t' = 0$, respectively



4-dimensional language it is a 2-dimensional surface (a world sheet, see Fig. 6.14) formed by the world lines of all points on the ruler. So now the length of a ruler, a concept crystal clear in the 3-dimensional language, would become vague in the 4-dimensional language: the length of which line segment is the length of the ruler? Those who speak the 4-dimensional language know that a ruler is not a 1-dimensional object; rather, it is 2-dimensional. This is an absolute object, which does not depend on reference frames, coordinate systems, or observers. Why is a ruler a 1-dimensional object in the traditional (3-dimensional) language? This is because people see things from the perspective of their own reference frames, which are relative in the first place. Since each surface Σ_t of simultaneity of an inertial frame \mathcal{R} represents the whole space at t , the intersection of Σ_t and the ruler naturally represents “the ruler measured (seen) by an observer in the frame \mathcal{R} at t ”. For instance, the intersection oa of the surface of simultaneity at $t = 0$ and the world sheet of the ruler is the ruler measured by \mathcal{R} at $t = 0$. Suppose the difference of the spatial coordinates between o and a are Δx , Δy and Δz , then the length of the ruler measured by \mathcal{R} is $(\Delta x^2 + \Delta y^2 + \Delta z^2)^{1/2}$, i.e., the length of the line segment oa . [One can either say it is the Euclidean arc length, as oa is a line lying on a 3-dimensional Euclidean space (the surface of simultaneity), or one can say it is the Minkowski arc length, as oa is a line in 4-dimensional Minkowski spacetime. The fact that t is a constant on the surface of simultaneity assures the consistency of these two viewpoints]. However, the simultaneity is relative, and the intersection (line segment ob) of the world sheet of the same ruler and the surface of simultaneity of a frame \mathcal{R}' at $t' = 0$ represents “the ruler measured by an observer in the frame \mathcal{R}' at $t' = 0$ ”; hence, the length of the ruler should be the length of ob . Since oa and ob are two different line segments in spacetime, it is not surprising that they have different lengths. Therefore, length contraction (also called Lorentz contraction) is not caused by elasticity or any physical mechanism (there is no “contraction” at all), and the essential cause of it is nothing but the fact that different surfaces of simultaneity for different reference frames lead to different measurements of 1-dimensional rulers (which are relative), although there is only one ruler (only one world sheet of the ruler). So it is certainly

Fig. 6.15 Based on the simultaneity lines $t = t_b$ and $t = 0$ of \mathcal{R} , this reference frame regards that C' is the clock which runs slower



not surprising that different 1-dimensional rulers have different lengths. In this way, “length contraction” is just like the classic parable of the blind men and an elephant.¹

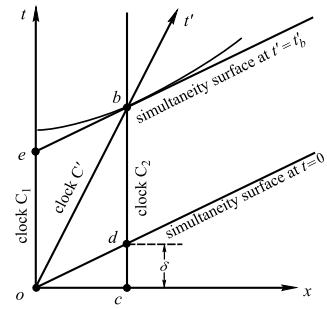
Since the frame \mathcal{R} treats the ruler as at rest while \mathcal{R}' treats the ruler as moving, the lengths of the line segments oa and ob are the rest length and “moving” length, respectively. Now the only problem left is to compare l_{oa} and l_{ob} . Intuitively we have $l_{ob} > l_{oa}$, so it seems that the moving ruler is longer! However, this is the “deception” of the spacetime diagram again. From the calibration curve passing through a we see that $l_{ob} < l_{oa}$, and thus the moving ruler is shorter. To find the quantitative relation between them, we just need to compute the length of both line segments. The arc length of a spacetime curve is an absolute quantity and is independent of the coordinate system. For the convenience of comparison, we choose the same coordinate system (the inertial frame that corresponds to \mathcal{R}) to compute both of them. Noticing that the coordinates of o in this system are $(0, 0, 0, 0)$, from the expression of the Minkowski line element in this system we obtain $l_{oa} = \sqrt{x_a^2 - 0} = x_a$, and $l_{ob} = \sqrt{x_b^2 - t_b^2}$. Also, it follows from (6.1.5) that the equation for the x' -axis is $t = vx$, and hence $t_b = vx_b$. From Fig. 6.14 we can see that $x_b = x_a$, and plugging these into the equations above yields $l_{ob} = \gamma^{-1}x_b = \gamma^{-1}x_a = \gamma^{-1}l_{oa}$. This is the well-known quantitative relation for length contraction.

6.2.2 Time Dilation

Consider two standard clocks C_1 and C_2 in an inertial frame \mathcal{R} and a standard clock C' in another inertial frame \mathcal{R}' . The world lines of these three clocks are shown in Fig. 6.15. From the viewpoint of \mathcal{R} , the clocks C_1 and C_2 are at rest, while C' is moving. At the beginning, C' and C coincide at event o , at which both clocks are zeroed. After a while, C' will coincide with C_2 at event b . From the fact that proper

¹ This is the story of a group of blind men who have never come across an elephant trying to conceptualize the elephant by touching it; however, their understandings of an elephant turn out to be completely different, since each of them touches a different part of the elephant’s body.

Fig. 6.16 Based on the simultaneity lines $t' = t'_b$ and $t' = 0$ of \mathcal{R}' , this frame regards that C_2 is the clock which runs slower



time is equal to the arc length of the world line, we can see that the reading of the clock C' at b is equal to l_{ob} . Both C_2 and C_1 belong to the same frame \mathcal{R} , and the x -axis is a line of simultaneity of \mathcal{R} , so since the reading of C_1 at o is zero, the reading of C_2 at c should also be zero. Hence, the reading of C_2 at b equals $l_{cb} = l_{oa}$. Plotting the calibration curve through p we see that $l_{ob} < l_{oa} = l_{cb}$, and thus the frame \mathcal{R} regards C' (the moving clock) as running slower. However, from the viewpoint of \mathcal{R}' , the event o happens simultaneously with d (see Fig. 6.16) rather than c . Since the reading of C_2 is zero at c , it must have a reading $\delta > 0$ at d . A short time later, C_2 coincides with C' at b . Although the reading l_{ob} of C' is smaller than the reading l_{cb} of C_2 (which is admitted by both frames), the observer in \mathcal{R}' does not admit that C' runs slower, because when C' is zeroed, judged by the line of simultaneity of \mathcal{R}' , C_2 has already had a reading δ (C_2 “jumped the gun”). Hence, one should subtract δ from the reading l_{cb} of C_2 at b , and then compare with l_{ob} , i.e., the observer in \mathcal{R}' thinks that one should compare l_{db} and l_{ob} . From the calibration curve passing through b we know that $l_{ob} > l_{oe} = l_{db}$, and hence \mathcal{R}' regards C_2 as running slower, and it is still the moving clock that runs slower. Figure 6.17 is a 3-dimensional demonstration of the discussion above, where (a) and (b) are the 3-dimensional perspectives of \mathcal{R} and \mathcal{R}' , respectively. Thus, we can see again that the 3-dimensional perspective depends on the reference frame; only spacetime diagrams and the 4-dimensional formulation can be independent of the reference frame.

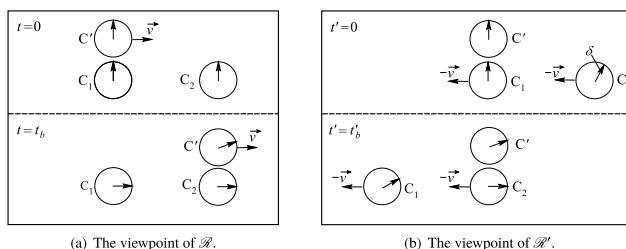


Fig. 6.17 The 3-dimensional perspectives of \mathcal{R} and \mathcal{R}'

Following the derivation of the quantitative relation between the lengths of a rest and a moving ruler, and noticing $x_b = vt_b$, it is not difficult to find from Fig. 6.15 that the quantitative relation between the time intervals of a rest and a moving clock is $l_{ob} = \gamma^{-1} l_{oa}$.

The discussion above clearly indicates that, just like nothing is really contracting in the phenomenon of “length contraction”, none of the clocks really runs slower in the phenomenon of “time dilation”. (They all have the rate of a standard clock, i.e., the difference of the readings equals the arc length). It should be emphasized that there are a variety of methods for “clock comparison” (comparing the readings of standard clocks in different inertial frames), which can lead to different results. Therefore, when talking about clock comparison, one must stipulate all the details of the method beforehand. The method we just used in the phenomenon of “time dilation” is standard, but it is only one of the various methods for clock comparison. A feature of this method is that it involves three clocks C_1 , C_2 and C' , two of which are in the same inertial frame and have been synchronized beforehand. Without C_2 , one can still get $l_{ob} < l_{oa}$ from Fig. 6.15; however, one cannot conclude that “the observer who carries C_1 measures that C_2 runs slower”, because there is no way for that observer to measure it directly since the event b is not on the world line of C_1 . The only way for C_1 to observe b is to receive a light signal (or other signals) coming from b , which leads to the problem that the propagation of light takes time. (One can certainly apply this method, but to do so this problem needs to be taken into account). Actually, when we draw the conclusion that “ C' is slower” by means of C_2 , we have already cleverly let the light signal play the role of a “messenger”, since a light signal has been used when synchronizing C_1 and C_2 (see Sect. 6.1.4). In summary, without C_2 , one cannot compare C_1 and C' using the method above; in other words, this method of clock comparison will not have any physical meaning without the third clock.

When there are only two clocks C and C' , there are still methods of clock comparison that are physically meaningful. For example, as shown in Fig. 6.18, the observer G that carries the clock C can compare the clocks by looking at two clocks using the left and right eyes respectively at a time a . “Looking at C' using the right eye” means that the light signal sent from C' at e is received by the right eye at a (the photon goes from e to a through a future-directed null geodesic). If both clocks are zeroed at o , then the reading of C at a equals l_{oa} , and the reading of C' at e equals l_{oe} . Since $l_{oe} < l_{oa}$, the observer G will also conclude that “the moving clock runs slower”, but the difference is that this method will make the moving clock even slower when compared with the method in Fig. 6.15. To compute how much the clock slows down, one can make a parallel line passing through e and intersect with the world line of C at f (see Fig. 6.19). Let $\tau \equiv l_{oa}$, $\tau' \equiv l_{oe}$, $p = l_{of}$, $q \equiv l_{fa}$, then $l_{ef} = q$. From $p = \gamma \tau'$ (the quantitative relation for the regular time dilation) and $u = q/p$, the geometric expression for the relative speed of the two clocks (C regards C' as moved for a distance q during a time p), we can easily find that²

² This is the relativistic Doppler relation. It is valid for both positive and negative u , giving respectively red and blue shift (see Sect. 6.6.6).

Fig. 6.18 G looks at C from the left eye and C' from the right eye at a , and finds that the moving clock C' runs slower. The observation is a “redshift” since the clock is moving away

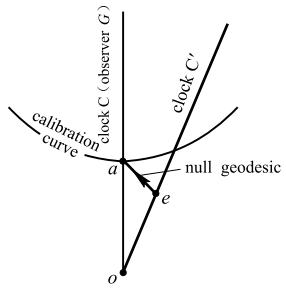


Fig. 6.19 The simple geometric method for finding the relation between τ and τ'

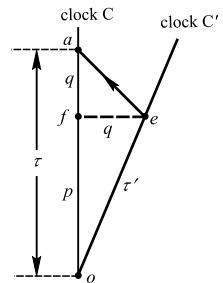
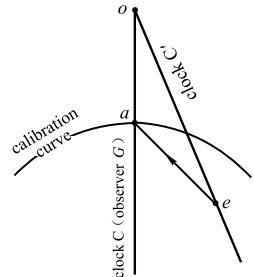


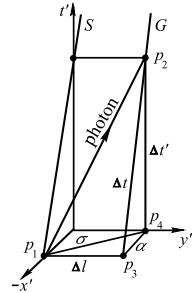
Fig. 6.20 G looks at C from the left eye and C' from the right eye at a , and finds that the moving clock C' runs faster. The observation is a “blueshift” since the clock is moving towards the observer



$$\tau' = \sqrt{(1-u)/(1+u)}\tau . \quad (6.2.1)$$

One can even come up with a method of clock comparison (see Fig. 6.20) that leads to the result that “the moving clock runs faster”! Suppose C and C' are both zeroed at o , then an observer G who uses two eyes to observe C and C' respectively at a will get negative readings from both clocks. From the figure it is easy to see that $l_{oa} < l_{oe}$, and hence the reading of C' is even more negative than that of C . Thus, G will think “the moving clock runs faster”. Though it sounds ridiculous, this conclusion is beyond reproach: it is nothing but a result coming from the specific method of clock comparison in Fig. 6.20. Therefore, to compare clocks we must indicate all the details of the method beforehand, for which a spacetime diagram would be very helpful.

Fig. 6.21 The spacetime diagram for Example 1



The examples above only involve 1-dimensional space. Here is an example in 2-dimensional space [Guo (2008) p. 235, Problem 5].

Example 1 A light source S and a receiver G are at rest in an inertial frame \mathcal{R} . The distance between them is Δl . Immerse the $S - G$ apparatus in a homogeneous infinite liquid medium (whose rest refractive index is n). Suppose the liquid is flowing with a speed u relative to \mathcal{R} along a direction that is perpendicular to the line connecting S and G . Find the time Δt for a light signal going from S to G (measured in \mathcal{R}).

Solution Denote the inertial frame at rest relative to the liquid as \mathcal{R}' . Draw a 3-dimensional spacetime diagram based on the frame \mathcal{R}' (see Fig. 6.21), where events p_1 and p_2 represent a photon coming out of S and arriving at G , respectively, and hence the line segment $p_1 p_2$ represents the propagation of that photon. Viewed from \mathcal{R} and \mathcal{R}' , the times this propagation takes are, respectively, the Minkowski length of the line segments $p_3 p_2$ and $p_4 p_2$ (the world line S can be used as the t -axis for the frame \mathcal{R}). The length of $p_1 p_3$ is obviously Δl . Using σ and α to represent the length of $p_1 p_4$ and $p_3 p_4$, we can easily see that

$$\Delta t = \gamma^{-1} \Delta t', \quad \text{where } \gamma \equiv (1 - u^2)^{-1/2} \quad (\text{the phenomenon of "time dilation"},)$$

$$u = \frac{\alpha}{\Delta t'} \quad (\text{in } \mathcal{R}', G \text{ moved for a distance } \alpha \text{ in } \Delta t' \text{ with a speed } u),$$

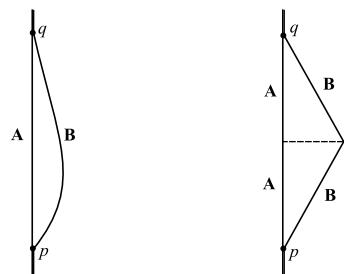
$$\sigma^2 = (\Delta l)^2 + \alpha^2 \quad (\text{the Pythagorean theorem in 3-dimensional Euclidean space}),$$

$$\frac{1}{n} = \frac{\sigma}{\Delta t'} \quad (\text{the speed of light is isotropic in a medium at rest, whose value is } 1/n).$$

Solving the equations above yields

$$\Delta t = \sqrt{\frac{1 - u^2}{n^2 - u^2}} \Delta l.$$
■

Fig. 6.22 The twin “paradox”



(a) Spacetime diagram.

(b) The simplest example.

6.2.3 The Twin “Paradox”

Figure 6.22a is the spacetime diagram for the twin “paradox”, where the two curves are the world lines of the twin brothers A and B . The curve A is a vertical line indicating that A stays at home (as an inertial observer), while the curve B is a non-geodesic indicating that B goes for a journey in space and returns. Suppose the brothers are of the same age when they separate, would they be the same age when they meet again? If not, then who is older? This is nothing but a question of comparing the proper times of A and B between p and q , i.e., a question of comparing the arc lengths l_A and l_B between p and q . Since the timelike geodesic is the longest timelike curve between two points in Minkowski spacetime (see the paragraph above Optional Reading 3.3), we have $l_A > l_B$, and thus B is younger than A when they meet again. Figure 6.22b is the simplest example of the twin “paradox” (where the world line of B is a broken line composed by two timelike geodesics); using the quantitative relation for time dilation, it is easy to see that $l_A = \gamma l_B > l_B$.

These are the essentials of the twin “paradox”, and the problem itself is just that simple. However, due to a lack of deep understanding at the early stage of relativity study, people used to consider this problem as a paradox. The argument regarding the twin “paradox” even had an upsurge in 1957–1958 (though most physicists had agreed that the problem had been solved long ago), and some papers were even published in journals like *Nature* and *Science*. The representatives for the two sides were physicist W. H. McCrea and physicist and natural philosopher H. Dingle. Dingle claimed that according to relativity everything is relative, and thus the twins should be the same age when they meet again. McCrea, however, pointed out shrewdly that it is not true that everything is relative in relativity; it is the fact that the twin brother B has an acceleration while A does not have one which leads to the result of the age difference. As the study went deeper, especially after the geometric language became widely used, physicists have already reached a consensus on the twin “paradox”, which is exactly what we have shown at the beginning of this subsection [see, for example, Sachs and Wu (1977) p. 42–43; Wald (1977) pp. 25–26; Misner et al. (1973) p. 167]. It should be particularly emphasized that one should not have the idea that “everything is relative in relativity” based solely on the name of the theory, as this is a critical misunderstanding!

The twin “paradox” was verified experimentally in 1971 using cesium atomic clocks, not humans, of course; the reader may refer to Hafele and Keating (1972a; 1972b) for more information and look at Exercise 6.10.

Now, we answer a few frequently asked question regarding the twin “paradox”.

Q: In the phenomenon of time dilation, the two observers are on an equal footing: *A* thinks *B*’s clock runs slower, and *B* thinks *A*’s clock runs slower. Why in the twin “paradox” are *A* and *B* not on an equal footing (everyone thinks *B* is younger than *A*)?

A: The premise for these two phenomena are different. In the phenomenon of time dilation, both observers are undergoing inertial motion; since inertial frames are on an equal footing, the result for both of them are certainly the same. However, in the twin “paradox”, one of the brothers is experiencing a non-inertial motion (the world line is not a geodesic), otherwise they will not meet again after they have separated. The premise implies an inequality between the two observers, and thus the conclusion is also one-sided.

Q: The conclusion of the twin “paradox” is that the accelerating brother is younger. However, since acceleration is relative, *B* accelerating relative to *A* means that *A* is accelerating relative to *B*. In this way, wouldn’t *B* also think *A* is younger?

A: One needs to distinguish the 3-dimensional and 4-dimensional accelerations (namely 3-acceleration and 4-acceleration, see Sect. 6.3), the former of which is relative, while the latter of which is absolute (independent of the choice of the observer, reference frame, coordinate system, etc.). On the other hand, the concept of inertial motion and non-inertial motion are both absolute: a point mass undergoes inertial motion if and only if its world line is a geodesic (independent of the reference frame!). When we consider the term “accelerated motion” as a synonym of “non-inertial motion”, it is supposed to be understood as the 4-acceleration. If one uses the word “acceleration” to describe the twin “paradox”, then one should say “the brother with a 4-acceleration is younger”. No observer would say that *A* has a 4-acceleration, and there is no issue anymore. It is already a convention for physicists that in the 3-dimensional language, when talking about acceleration without specifying a reference frame, it is always assumed to be relative to an inertial frame. Under this agreement, expressions like “the brother experiencing accelerated motion is younger” and “an electric charge only radiates under accelerated motion” are both correct.

Q: It is sometimes heard that the twin “paradox” is inside the scope of general relativity, and thus cannot be interpreted by just using special relativity. Is that right?

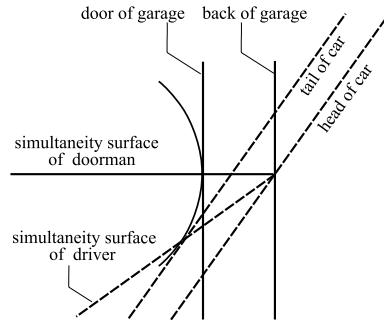
A: No. Didn’t we just interpret it clearly in the first paragraph of this subsection? The misconception of some people that the twin “paradox” is related to general relativity may occur when they choose the coordinate system corresponding to the reference frame of *B* for calculating the time experienced by *B*. This is a non-inertial frame, and some people may think general relativity is involved as long as we talk about non-inertial frames. The explanations of this misunderstanding are the following: ① The time experienced by *B* is the length of his world line, which is a geometric quantity independent of the coordinate system, and hence it is not necessary at all to trouble yourself by choosing a non-inertial frame for the

calculation. ② At the very least, even if you insist on using a non-inertial frame for the calculation, there is no need to use general relativity at all. We should specify the division criteria for special versus general relativity. At first, people thought they could use coordinate systems as the criterion, and it would be considered to be in the scope of general relativity as long as a non-inertial frame is involved. Later, however, people realized that it is much more natural (and elegant) to use the absolute spacetime geometry (which is independent of any human choice) as the criterion. Therefore, now we shall have the following standard: any physics problem that has Minkowski spacetime as its background is in the scope of special relativity, while general relativity must be used when spacetime is curved (see Chap. 7). When discussing any physics problem, an important but often ignored step is to identify the spacetime background beforehand, i.e., to specify in what spacetime the physical phenomenon happens. The premise for the twin “paradox” is that the whole process happens in Minkowski spacetime, and thus is in the realm of special relativity (unless one stipulates that the background spacetime is not Minkowski, which means the gravitational field is not negligible, see Chap. 7). Unfortunately, some people went even further and mistakenly thought that accelerated motion could lead to curved spacetime and so general relativity must be involved. (Some may conclude that the spacetime is curved just based on the fact that the Christoffel symbols $\Gamma^\mu_{\nu\sigma}$ do not all vanish in a non-inertial coordinate system, but do not realize that it is absolutely normal to find the $\Gamma^\mu_{\nu\sigma}$ of a Minkowski metric to be nonvanishing in a non-inertial frame). Another famous example similar to this is Einstein’s rotating disk, which is sometimes also misunderstood as a problem involving general relativity. The premise of this problem is actually also that the whole phenomenon (including the motions of the disk and the observers on it) takes place in Minkowski spacetime, and therefore it is also within the scope of special relativity. The best way to clearly analyze the problem of Einstein’s rotating disk is also by using the geometric language; however, it is much more complex than the twin “paradox”, see Sect. 14.2 (Volume II) for details.

6.2.4 The Garage “Paradox”

Suppose a car has the same rest length as a garage. When driving the car into the garage, the driver thinks: “the moving garage becomes shorter, which will not be large enough to fit the car.” However, the doorman of the garage thinks: “the moving car has shrunk, and the garage will be more than enough to fit the car.” Is the driver correct? Is the doorman correct? It will be crystal clear in the 4-dimensional geometric language. To make the problem simple and more specific, we assume that the garage does not have a back wall (its “back wall” will be just a line on the ground). Figure 6.23 is the spacetime diagram of the car coming into the garage at a uniform speed (one can use a calibration curve to make sure the car and the garage have the same rest length in the diagram). It is easy to see from the diagram that, measured by the inertial frame of the doorman, the garage is longer than the car and has enough room for the car;

Fig. 6.23 The spacetime diagram for the garage “paradox” (without a real back wall)



measured by the inertial frame of the driver, however, the car is longer than the garage and cannot be fit into the garage. The viewpoints of them are both correct, and the divergence of their conclusion comes from the relativity of simultaneity. Actually, a question like “can the car be fit in or not?” is not well-defined and should not be raised. Since the conclusion is relative, this absolute type of question is meaningless, just like in the phenomenon of “length contraction” one cannot ask “which ruler is longer?” The case where the garage has a hard back wall is a little more complicated, the basic principle is: due to relativity, any information cannot propagate faster than the speed of light, and so the information that the head of the car crashes into wall (and stops moving) takes time to propagate to the tail of the car; the tail of the car will start to decelerate and come to a stop after “receiving” this information. Therefore, the car would be physically compressed into a length that can be fit into the garage (no matter from whose perspective). Motivated readers should draw a spacetime diagram that roughly describes the whole process and finish Exercise 6.11.

6.3 Kinematics and Dynamics of a Point Mass

Due to the importance and the subtleties of concepts like momentum, energy and mass in special relativity, it is necessary to first review some relevant issues.

The principle of relativity requires that the laws of physics have the same form in all inertial frames. The transformation between reference frames in Newtonian mechanics is a Galilean transformation, while in special relativity it is a Lorentz transformation. Hence, in Newtonian mechanics the principle of relativity requires the mathematical expressions for the laws of physics to be invariant under Galilean transformations (called **Galilean covariance**), while in special relativity it requires the mathematical expressions for the laws of physics to be invariant under Lorentz transformations (called **Lorentz covariance**). This principle is very powerful in that it is a “law of laws”, which means any law that does not have Lorentz covariance must be modified in order to be fit into special relativity. One notable example is the law of conservation of momentum. In Newtonian mechanics, the momentum of a point mass is defined as the product of the mass m and the velocity \vec{u} , i.e., $\vec{p} := m\vec{u}$, and the corresponding force is defined by the time rate of change of the

Fig. 6.24 A perfectly inelastic collision of two identical balls

frame	before		after
\mathcal{R}'	1 ● → \vec{v}	2 ● ← $-\vec{v}$	● at rest
\mathcal{R}	1 ● → \vec{u}	2 ● at rest	● → \vec{v}

point mass's momentum, i.e., $\vec{f} := d\vec{p}/dt$; plugging in the definition of \vec{p} yields $\vec{f} = m d\vec{u}/dt = m \vec{a}$. Thus, even though $\vec{f} = m \vec{a}$ is called Newton's second law, it is just a definition of a force; only when combined with the expression for a force under a specific physics circumstance can it provide a law of physics. For instance, in the circumstance of a spring, combined with $\vec{f} = -K \vec{x}$ it yields $d\vec{p}/dt = -K \vec{x}$, which is Hooke's law. Now let us consider the collision of two balls in terms of the principle of relativity. Use \vec{p}_1 and \vec{f}_1 to represent, respectively, the momentum of ball 1 and the force acting on it (from ball 2), then $\vec{f}_1 = d\vec{p}_1/dt$, and similarly $\vec{f}_2 = d\vec{p}_2/dt$. Newton's third law guarantees that $\vec{f}_1 = -\vec{f}_2$, and hence $d(\vec{p}_1 + \vec{p}_2)/dt = 0$, i.e., momentum is conserved in the collision. Thus, the conservation of momentum is the result of combining the definition of a force with Newton's third law. If we observe the same collision process from another inertial frame, then according to the velocity transformation formula derived from the Galilean transformation it is not difficult to see that the momentum is still conserved, and so the conservation of momentum is Galilean covariant, which satisfies the principle of relativity. However, if one still uses the Newtonian definition of momentum in special relativity, i.e., $\vec{p} := m \vec{u}$ ($m = \text{constant}$), then the following simple example is sufficient to show that the conservation of this momentum is not Lorentz covariant. Consider a perfectly inelastic collision of two identical balls. Suppose the velocities of these two balls in the frame \mathcal{R}' are equal in magnitude and opposite in direction before the collision (and thus the total momentum is zero), then from the symmetry we can see that the velocities of both balls will be zero (see Fig. 6.24), which indicates that the momentum is conserved in \mathcal{R}' in the collision process. Now we observe this process in the frame \mathcal{R} . Suppose ball 2 is at rest relative to \mathcal{R} , then the velocity of \mathcal{R}' relative to \mathcal{R} is equal to the velocity \vec{v} of ball 1 relative to \mathcal{R}' before the collision. From the relativistic velocity transformation formula (which can be found in any textbook on special relativity) we know that the velocity of ball 1 in the frame \mathcal{R} is (for now we keep the speed of light explicit, rather than set $c = 1$)

$$u = \frac{v + v}{1 + v^2/c^2} = \frac{2v}{1 + v^2/c^2}. \quad (6.3.1)$$

Suppose the Newtonian mass for both of the balls is m , then the total momenta of the balls in \mathcal{R} before and after the collision are, respectively,

$$\text{initial total momentum (magnitude)} = mu + 0 = \frac{2mv}{1 + v^2/c^2},$$

$$\text{final total momentum (magnitude)} = 2mv.$$

(The conservation of Newtonian mass is used in the second line). The total momenta before and after the collision are not the same, and thus the momentum is not conserved in the frame \mathcal{R} . This indicates that now the conservation of momentum is not Lorentz covariant, and hence is not a law. Now we have two choices: either give up on the conservation of momentum or render the conservation of momentum Lorentz covariant by modifying the definitions of mass and momentum. In consideration of the significance of the laws of conservation in physics, we certainly would like to choose the latter one. To get an idea how to modify them, let us consider the following: suppose a point mass is accelerated by a constant force. Based on Newton's second law, its speed must eventually exceed the speed of light at some point in time, which contradicts special relativity. In order to get rid of this inconsistency, it would be reasonable to suspect that the mass of a point mass increases with its speed in relativity. In this way, the acceleration of the point mass under a constant force would become smaller and smaller, and so it is possible that its speed will never reach the speed of light. Therefore, we can suggest the following modification: we still define momentum as mass times velocity; however, the mass, now denoted by m_u (called the **relativistic mass**), is no longer a constant but depends on the speed u . Now based on this idea let us reconsider the conservation of momentum in the frame \mathcal{R} in Fig. 6.24. Since ball 2 is at rest before the collision while ball 1 moves with velocity u , their relativistic masses are, respectively, m_0 (called the **rest mass**) and m_u . Hence,

$$\text{initial total momentum (magnitude)} = m_u u + 0 = \frac{2m_u v}{1 + v^2/c^2}, \quad (6.3.2)$$

$$\text{final total momentum (magnitude)} = M_v v, \quad (6.3.3)$$

where M_v represents the total mass of the combined body after the collision. Assume that the total mass is invariant in the collision, i.e., $m_u + m_0 = M_v$. (This is a very natural assumption, the meaning of which will be explained later). Then (6.3.3) becomes

$$\text{final total momentum (magnitude)} = (m_u + m_0)v. \quad (6.3.4)$$

Comparing (6.3.2) and (6.3.4) we can see that, in order to let the conservation of momentum hold in the frame \mathcal{R} , we only have to require that

$$m_u = m_0 \frac{1 + v^2/c^2}{1 - v^2/c^2}. \quad (6.3.5)$$

Also, a simple calculation starting from (6.3.1) shows that

$$\sqrt{1 - u^2/c^2} = \frac{1 - v^2/c^2}{1 + v^2/c^2}, \quad (6.3.6)$$

and comparing this with (6.3.5) yields

$$m_u = \frac{m_0}{\sqrt{1 - u^2/c^2}}. \quad (6.3.7)$$

Thus, for the collision shown in Fig. 6.24, we can only guarantee that momentum is conserved in \mathcal{R} if we allow m_u to change with the speed u according to (6.3.7). From this, we extrapolate that the momentum of a point mass in special relativity should be defined as

$$\vec{p} := m_u \vec{u} \quad [\text{where } m_u \text{ is given by (6.3.7)}]. \quad (6.3.8)$$

Usually we denote $\gamma_u \equiv (1 - u^2/c^2)^{-1/2}$, and hence the momentum can also be expressed as

$$\vec{p} = \gamma_u m_0 \vec{u}, \quad \text{or, for short, } \vec{p} = \gamma m_0 \vec{u} = m_u \vec{u}. \quad (6.3.9)$$

Now that we have the definition of momentum, we can define force. In relativity, the force \vec{f} acting on a point mass is still defined by the time rate of change of the point mass's momentum:

$$\vec{f} := \frac{d\vec{p}}{dt}. \quad (6.3.10)$$

The principle of relativity requires the above equation to be Lorentz covariant, which determines the transformation law of forces between inertial frames (see textbooks on special relativity for details).

Now we introduce the definition of energy. First, following Newtonian mechanics, we define the kinetic energy E_k of a point mass using the following two requirements: ① $E_k = 0$ when the point mass is at rest ($u = 0$), ② the time rate of change of the kinetic energy equals the power $\vec{f} \cdot \vec{u}$, from which we obtain

$$\frac{dE_k}{dt} = \vec{f} \cdot \vec{u} = \frac{d\vec{p}}{dt} \cdot \vec{u} = \vec{u} \cdot \frac{d(m_u \vec{u})}{dt} = m_u \vec{u} \cdot \frac{d\vec{u}}{dt} + \vec{u} \cdot \vec{u} \frac{dm_u}{dt} = m_u u \frac{du}{dt} + u^2 \frac{dm_u}{dt}, \quad (6.3.11)$$

where dm_u/dt can also be expressed using (6.3.1) as

$$\frac{dm_u}{dt} = \frac{d}{dt} \left(\frac{cm_0}{\sqrt{c^2 - u^2}} \right) = \frac{m_u u}{c^2 - u^2} \frac{du}{dt}. \quad (6.3.12)$$

Plugging this into (6.3.11) yields

$$\frac{dE_k}{dt} = (c^2 - u^2) \frac{dm_u}{dt} + u^2 \frac{dm_u}{dt} = c^2 \frac{dm_u}{dt}. \quad (6.3.13)$$

Noticing that $m_u = m_0$ and $E_k = 0$ when $u = 0$, by integrating over the above equation we find the kinetic energy at the speed u is

$$E_k(u) = c^2 \int_{m_0}^{m_u} dm = m_u c^2 - m_0 c^2. \quad (6.3.14)$$

Albert Einstein boldly claimed that the $m_u c^2$ on the right-hand side of the equation above to be the (total) energy of the point mass at the speed u (denoted by $E = mc^2$, where m is short for m_u). Thus, $m_0 c^2$ is the mass when the point mass is at rest (denoted by $E_0 = m_0 c^2$, called the **rest energy** of the point mass), and the kinetic energy is the difference of the total energy and rest energy. $E = mc^2$ indicates that the energy E is proportional to the mass m (by which we mean the relativistic mass m_u), called the equivalence of mass and energy. In the geometrized unit system $c = 1$, and hence $E = m$, i.e., energy is equal to mass, and $E_0 = m_0$ indicates that an object has the same amount of energy as its rest mass even at rest. This is an incredibly huge amount of energy: the energy of an object with $m_0 = 1 \text{ g}$ (which is about 1% of a bag of instant noodles) is

$$m_0 c^2 = 10^{-3} \times (3 \times 10^8)^2 = 9 \times 10^{13} \text{ J},$$

which is roughly the energy released by an atomic bomb!

In Newtonian mechanics, there are both the law of conservation of mass and the law of conservation of energy. What about in special relativity? First of all, the energy defined by $E = mc^2$ (NB: m is short for m_u) satisfies the law of conservation of energy. This fact should be regarded as a theoretical hypothesis, which has been supported by numerous experiments. As for whether the mass is conserved, it depends on which mass you are talking about. Since $E = mc^2$, the conservation of E also implies the conservation of the relativistic mass m , and they are not independent.³ As to the rest mass, we should emphasize that it does not obey the conservation law. For instance, suppose in a fission process an atomic nucleus at rest is split into two pieces (both are moving). Use M, m_1, m_2 to represent the relativistic mass of the the nucleus and the two pieces, respectively, then from energy conservation we have

$$Mc^2 = m_1 c^2 + m_2 c^2. \quad (6.3.15)$$

Before the fission the nucleus is at rest, the relativistic mass of which is equal to the rest mass M_0 . Use m_{01}, m_{02}, u_1 and u_2 to respectively represent the rest masses and the velocities of the two pieces. Let $\gamma_1 \equiv (1 - u_1^2/c^2)^{-1/2}$, $\gamma_2 \equiv (1 - u_2^2/c^2)^{-1/2}$, then $m_1 = \gamma_1 m_{01}$, $m_2 = \gamma_2 m_{02}$. Hence, it follows from (6.3.15) that

$$M_0 = \gamma_1 m_{01} + \gamma_2 m_{02} > m_{01} + m_{02}. \quad (6.3.16)$$

Thus, the rest mass is not conserved! The difference $\Delta m_0 = M_0 - (m_{01} + m_{02})$ is called the **mass defect**. In summary: in special relativity there are in total only two laws of conservation regarding momentum, energy, rest mass and relativistic mass,

³ However, make sure not to think this “law of the conservation of mass” as the same as the one in Newtonian mechanics. The former is a conservation law of a physical quantity (the relativistic mass), while the latter reflects the following tenet of Newton: matter can neither be created nor destroyed. From today’s vantage point, this tenet is not quite true, since matter can be “destroyed”—it can be turned into radiation, even though the energy does not change. Thus, energy is conserved while matter is not.

namely the conservation of momentum and the conservation of energy. Before, we used to assume $m_u + m_0 = M_v$ when we rewrote (6.3.3) into (6.3.4), and now we see that this equation represents the conservation of energy. Thus, energy should be assumed to be conserved when proving the Lorentz covariance of the momentum $\vec{p} = \gamma m_0 \vec{u}$.

In the original formulation of special relativity all four of these concepts existed: rest mass m_0 , rest energy E_0 , relativistic mass m (i.e., m_u) and total energy E . However, the relations $E = mc^2$ and $E_0 = m_0c^2$ indicate that among them there are only two independent ones. In fact, the modern literature (except for popular science) usually only keeps two concepts: mass and energy, where “mass” m refers to the rest mass (since we only keep one mass, there is no need to add “rest” to it, and the subscript “0” is also unnecessary), and “energy” refers to the total energy E . Now the relationship between E and m is $E = \gamma mc^2$ [where $\gamma \equiv (1 - u^2/c^2)^{-1/2}$]. In this way, there is only the law of the conservation of energy, while there is no law of the conservation of mass (notice that there is a mass defect). Later on, when we talk about mass in this text, unless otherwise stated, we always refer to the rest mass, and denote it by m (although we have used m for the relativistic mass before). Since the geometrized unit system is adopted, we have $E = \gamma m$. Having experienced a winding development course in the early days of special relativity, Einstein wrote in 1948 in a personal letter: “It is not good to introduce the mass $M = m(1 - v^2/c^2)^{-1/2}$ of a body for which no clear definition can be given, It is better to introduce no other mass concept other than the ‘rest mass’ m .”

The above is a review. From now on we go back to the geometrized unit system, in which $c = 1$. Before we introduce the 4-dimensional formulation of particle kinematics and dynamics, it is necessary to lay out the major definitions and relations in the 3-dimensional formulation as follows (except for the mass m and charge q which does not depend on the observer, all the quantities are defined relative to the inertial frame $\{t, x, y, z\}$):

$$\text{3-velocity (short for 3 – dimensional velocity) of a point mass } \vec{u} := d\vec{r}/dt, \quad (6.3.17)$$

where the position vector $\vec{r} = \vec{i}x + \vec{j}y + \vec{k}z$.

$$\text{3-acceleration of a point mass } \vec{a} := d\vec{u}/dt. \quad (6.3.18)$$

$$\text{3-momentum of a point mass } \vec{p} := \gamma m \vec{u}, \quad \gamma \equiv (1 - u^2)^{-1/2}, \quad u \equiv |\vec{u}|. \quad (6.3.19)$$

$$\text{Energy of a point mass } E := \gamma m. \quad (6.3.20)$$

$$\text{3-force acting on a point mass } \vec{f} := d\vec{p}/dt. \quad (6.3.21)$$

$$\text{The relation between the power of the 3 – force } \vec{f} \text{ and the energy of the point mass } \vec{f} \cdot \vec{u} := dE/dt. \quad (6.3.22)$$

3-force acting on a charged point mass in an electromagnetic field (Lorentz force)

$$\vec{f} := q(\vec{E} + \vec{u} \times \vec{B}), \quad (6.3.23)$$

where q is the electric charge of the point mass, \vec{u} is the 3-velocity, \vec{E} and \vec{B} are the electric field and magnetic field, respectively.

Remark 1 ① The γ here is short for $\gamma_u \equiv (1 - u^2)^{-1/2}$, while the γ in the Lorentz transformation (6.1.5) stands for $(1 - v^2)^{-1/2}$, where v is the relative speed between two inertial frames, and u is the velocity of a particle with respect to the chosen inertial frame. ② The transformations of coordinate systems are frequently involved in relativity, and therefore people often use the term “invariant”. Note that “invariant” and “conserved quantity” are two different concepts. A **conserved quantity** is a quantity whose value remains a constant (does not change with time) in a physical process; an **invariant** is a quantity that does not change with human factors such as a coordinate system, reference frame, or observer. The former emphasizes the physical process, and the latter emphasizes the transformation of the coordinate system, etc. Energy is a conserved quantity rather than an invariant; (rest) mass is an invariant rather than a conserved quantity; the electric charge of a charged particle is both an invariant and a conserved quantity.

This is the 3-dimensional formulation based on a specific inertial frame. Now we will introduce the 4-dimensional formulation, as well as the relationship between the 3- and 4-dimensional languages.

Definition 1 The **4-dimensional velocity (4-velocity)** U^a of a point mass is the tangent vector of the world line (parametrized by the proper time τ) of the point mass, i.e.,

$$U^a := \left(\frac{\partial}{\partial \tau} \right)^a. \quad (6.3.24)$$

Proposition 6.3.1 Let $U_a \equiv \eta_{ab} U^b$, then $U^a U_a = -1$.

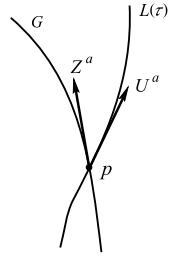
Proof The proper time is the arc length parameter of a timelike curve, and the tangent vector of a curve whose parameter is the arc length has unit length (see Sect. 2.5). \square

Remark 2 The 4-velocity is not defined outside the world line.

To observe the motion of a mass point, one can choose an arbitrary reference frame \mathcal{R} . Suppose $L(\tau)$ is the world line of the point mass, then for any point p on $L(\tau)$ there is always an observer G in \mathcal{R} passing through it (see Fig. 6.25), and so G can measure the event p . Let Z^a and U^a represent the 4-velocities of G and $L(\tau)$, respectively, at p . Physically, it is not difficult to understand that if $Z^a = U^a$, the observer G will think the point mass L is at rest at the time of p . Otherwise, the observer G will think the point mass has some kind of velocity (the 3-velocity) at the time of p . Before giving the definition of the 3-velocity of the point mass relative to the observer G at p , first we would like to set the stage as follows.

Imagine you are the observer G . ① You can only make direct measurements on the events happening on your world line. If an event happens outside your world line, you certainly may hear it or see it (observe indirectly), but it involves a signal propagating from this event to you, which takes some time and makes the

Fig. 6.25 The observer G and the point mass L intersect at p , so G can measure the event p



discussion a little complicated. (On the theoretical side, the shape of an object moving at high speed is a problem in this category; on the practical side, all the astronomy observation are indirect measurements). The simplest, clearest, and most basic kind of measurement is a direct measurement, i.e., the measurement of an event happening on the observer's world line, also called a **local measurement**. Luckily, a reference frame is formed by ubiquitous observers, and the events happening elsewhere can just be measured by another observer. ② When you measure an event happening at a point p on your world line, in many cases what is important is just the 4-velocity at p but not the whole world line. Then, there is no need to emphasize the world line of the observer, and one only needs to know the tangent vector Z^a of this world line at p . Hence, we can extract a more abstract concept, called an **instantaneous observer** [see Sachs and Wu (1977)], which contains two key elements, namely the point p and a (future-directed) timelike unit vector Z^a at p , together denoted by (p, Z^a) . ③ You, as an observer, have a sense of spatial direction besides a sense of time (from your standard clock). Assume you hold an arrow in your hand, and any direction it points to represents a spatial direction you can perceive. The collection of all the directions you can perceive at p (a point on your world line G) is of course a 3-dimensional set W_p , while for p as a point in \mathbb{R}^4 , its tangent space V_p is 4-dimensional. What is the relationship between W_p and V_p ? First let us consider the simplest case. Suppose you are an inertial observer in an inertial frame \mathcal{R} . The surface of simultaneity of \mathcal{R} is the 3-dimensional space of \mathcal{R} at a certain time, which is orthogonal to the world lines of all the inertial observers in this frame, and thus all the spatial vectors you have at p are orthogonal to your 4-velocity Z^a at p . Therefore, W_p corresponds to the 3-dimensional subspace of V_p orthogonal to Z^a , i.e.,

$$W_p = \{w^a \in V_p \mid \eta_{ab} w^a Z^b = 0\}.$$

This correspondence also applies to non-inertial observers, since we only care about the situation at one point p on the world line of the observer.

In Fig. 6.26, W_p is represented by as a small plane, but it is actually an “infinitesimal” plane. The most precise interpretation of W_p is a subspace of the tangent space at p , which in the figure can only be drawn as a small plane. Suppose $w^a \in V_p$, when $w^a \in W_p$, we say that w^a is a **spatial vector** for the observer G . A (nonzero) spatial vector must be a spacelike vector, but the converse is not true. From the definition

Fig. 6.26 W_p is the 3-dimensional subspace of V_p orthogonal to Z^a , any $w^a \in W_p$ can be seen as a spatial vector of G at p

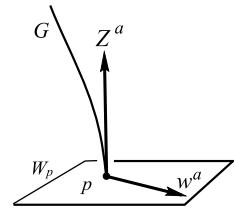
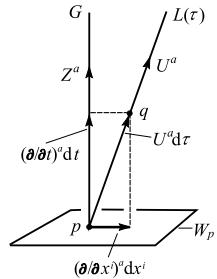


Fig. 6.27 The observer G measures that the spatial displacement of the point mass L during a time period dt is $(\partial/\partial x^i)^a dx^i$, and thus the 3-velocity should be defined by (6.3.27)



we can see that a spacelike vector is absolute (does not depend on factors such as the observer, reference frame or coordinate system, etc.), while a spatial vector is relative (depends on the 4-velocity Z^a of the observer). It follows from (4.4.2) that the induced metric of η_{ab} on W_p at p is $h_{ab} = \eta_{ab} + Z_a Z_b$, and from the paragraph below (4.4.4) we know that $h^a{}_b = \delta^a{}_b + Z^a Z_b$ is the projection map from V_p to W_p , i.e., $h^a{}_b v^b \in W_p$ is the projection of $v^a \in V_p$ onto W_p .

Suppose the world line L of a point mass and the world line G of an observer intersect at p , let us discuss the 3-velocity of L relative to G at p . First we discuss the case where $L(\tau)$ and G are both geodesics. Let U^a and Z^a be, respectively, the 4-velocities of $L(\tau)$ and G at p (see Fig. 6.27), and $\{t, x^i\}$ be the coordinates of the inertial frame that the inertial observer G belongs to. Then,

$$U^a = \left(\frac{\partial}{\partial \tau} \right)^a = \left(\frac{\partial}{\partial t} \right)^a \frac{dt}{d\tau} + \left(\frac{\partial}{\partial x^i} \right)^a \frac{dx^i}{d\tau}, \quad (6.3.25)$$

which can also be expressed as

$$U^a d\tau = \left(\frac{\partial}{\partial t} \right)^a dt + \left(\frac{\partial}{\partial x^i} \right)^a dx^i. \quad (6.3.26)$$

Suppose $p = L(\tau_1)$, and let $q \equiv L(\tau_1 + d\tau)$, then the geodesic segment pq represents the “infinitesimal” process of the point mass from the proper time τ_1 to $\tau_1 + d\tau$. For the observer G , the time of this process would be dt in (6.3.26), and the corresponding spatial displacement is $(\partial/\partial x^i)^a dx^i$. Hence, the **3-velocity of a point mass L relative to G** (also called the 3-velocity of L measured by G) should be defined as

$$u^a := \left(\frac{\partial}{\partial x^i} \right)^a \frac{dx^i}{dt} = \left(\frac{\partial}{\partial x^i} \right)^a \frac{dx^i/d\tau}{dt/d\tau}. \quad (6.3.27)$$

It follows from (6.3.25) that $\left(\frac{\partial}{\partial x^i} \right)^a \frac{dx^i}{d\tau}$ is the spatial projection of U^a , i.e., $h^a_b U^b$. Now if we set $\gamma \equiv dt/d\tau$, then (6.3.27) can be rewritten as

$$u^a := \frac{h^a_b U^b}{\gamma}. \quad (6.3.28)$$

The γ we just introduced (i.e., $\gamma \equiv dt/d\tau$) can also be expressed as

$$\gamma = -U^a Z_a, \quad (6.3.29)$$

since $-U^a Z_a = -\eta_{ab} U^a Z^b = -\eta_{\mu\nu} U^\mu (\partial/\partial t)^\nu = -\eta_{00} U^0 (\partial/\partial t)^0 = U^0 = dt/d\tau = \gamma$.

Remark 3 ① It is easy to see that the 3-velocity u^a is a spatial vector of the observer G at p (and thus can be denoted by \vec{u}). This is the most basic requirement for u^a : since the 3-velocity is a vector in the 3-dimensional language (called a 3-vector), of course it should be a spatial vector. ② Although we have used a coordinate system in the discussion above, the definition (6.3.28) of u^a is independent of the coordinate system. ③ Suppose \mathcal{R} is an inertial reference frame that the inertial observer G belongs to, then the u^a in (6.3.28) is also called the 3-velocity of the point mass L at p relative to \mathcal{R} . Suppose $\{t, x^i\}$ is an arbitrary inertial coordinate system in \mathcal{R} , then it follows from (6.3.27) that the components of the 3-velocity in this system are $u^i = dx^i/dt$. Note that the components of \vec{u} defined by (6.3.17) are also $u^i = dx^i/dt$, and thus this agrees with the definition of u^a in (6.3.28). ④ A 3-vector (e.g., u^a) at any point p in the 4-dimensional spacetime is an element in V_p , and hence is also a 4-vector, just the time component u^0 is zero.

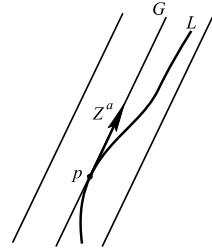
Since (6.3.28) only involves the tangent space of p (only involves an “infinitesimal” neighborhood of p), it also applies to the cases where $L(\tau)$ and G are not geodesics, and therefore we have the following definitions:

Definition 2 Suppose $L(\tau)$ is an arbitrary point mass, and $p \in L$, then the **3-velocity** u^a of the point mass relative to any instantaneous observer (p, Z^a) is defined by (6.3.28), where $h_{ab} = \eta_{ab} + Z_a Z_b$, and $\gamma \equiv -U^a Z_a$.

Definition 3 The magnitude $u = \sqrt{u^a u_a}$ of the 3-velocity vector u^a of a point mass with respect to an instantaneous observer is called the **3-speed** of the point mass with respect to this instantaneous observer, where $u_a := \eta_{ab} u^b = h_{ab} u^b$.

Remark 4 Suppose $p \in L$, and G is the geodesic determined by (p, Z^a) , then the 3-speed of the point mass L relative to the instantaneous observer (p, Z^a) agrees with the 3-speed of L relative to the inertial frame \mathcal{R} that G belongs to [defined by (6.1.2)]. For now we relax $L(\tau)$ to be either a timelike, null, or spacelike curve. For

Fig. 6.28 The instantaneous rest inertial reference frame of a point mass L at p



the timelike and spacelike cases, τ represents the arc length, and for the null case, τ represents an arbitrary parameter. Let $U^a \equiv (\partial/\partial\tau)^a$, we still use (6.3.28) to define u^a . Then,

$$\begin{aligned} u^2 &= h_{ab}u^a u^b = h_{ab}\frac{(h^a{}_c U^c)(h^b{}_d U^d)}{\gamma^2} = h_{cd}\frac{U^c U^d}{\gamma^2} \\ &= \frac{\eta_{cd}U^c U^d + Z_c Z_d U^c U^d}{\gamma^2} = \frac{\eta_{cd}U^c U^d + \gamma^2}{\gamma^2}. \end{aligned}$$

The equation above indicates that $u < 1 \Leftrightarrow \eta_{cd}U^c U^d < 0$, $u = 1 \Leftrightarrow \eta_{cd}U^c U^d = 0$, $u > 1 \Leftrightarrow \eta_{cd}U^c U^d > 0$. Thus, if we define the 3-velocity using (6.3.28), then the basic tenet of special relativity can be expressed using the 3-dimensional language as “the 3-speed of a point mass is slower than the speed of light”.

If the instantaneous observer (p, Z^a) is tangent to the world line L of the particle, then (p, Z^a) is called an **instantaneous rest observer** of this particle (the particle L is at rest at p to the observer). The geodesic G determined by p and Z^a is called the **instantaneous rest inertial observer** of L at p , and the inertial reference frame that G belongs to is called an **instantaneous rest inertial reference frame** of L at p , in which any inertial coordinate system is called an **instantaneous rest inertial coordinate system** of L at p . The concept of an instantaneous rest inertial frame will be very useful (Fig. 6.28).

Proposition 6.3.2 *The 4-velocity of a point mass can be 3 + 1-decomposed by means of an instantaneous observer (p, Z^a) :*

$$U^a = \gamma(Z^a + u^a), \quad (6.3.30)$$

where u^a is the 3-velocity of the point mass relative to the instantaneous observer, and $\gamma \equiv -Z^a U_a$.

Proof It follows from (6.3.28) that

$$\gamma u^a = h^a{}_b U^b = (\delta^a{}_b + Z^a Z_b) U^b = U^a - \gamma Z^a,$$

and hence we have (6.3.30). \square

Remark 5 From (6.3.30) we can see that γu^a is the spatial component of U^a . Choosing an inertial frame $\{t, x, y, z\}$ such that $(\partial/\partial t)^a = Z^a$, we can see from (6.3.30) that γZ^a is the time component of U^a . Hence, one can also write (6.3.30) as $U^a = \gamma(1, u^a)$, which agrees with the commonly used expression $U^\mu = \gamma(c, \vec{u})$ in texts on special relativity.

Remark 6 U^a is absolute (independent of any observer or coordinate system), while the $3 + 1$ decomposition of U^a depends on the observer (or coordinate system), and thus is relative. For another instantaneous observer (p, Z'^a) , the same U^a can be expressed as $U^a = \gamma' Z'^a + \gamma' u'^a$, i.e., both the time component $\gamma' Z'^a$ and the spatial component $\gamma' u'^a$ are different from γZ^a and γu^a .

Definition 4 Suppose the (rest) mass of a point mass is m , and the 4-velocity is U^a , then the **4-momentum** P^a of the point mass is defined as

$$P^a := mU^a. \quad (6.3.31)$$

Proposition 6.3.3 *The 4-momentum of a point mass can be $3 + 1$ -decomposed by means of an instantaneous observer (p, Z^a) :*

$$P^a = EZ^a + p^a, \quad (6.3.32)$$

where the energy E and the 3-momentum p^a are defined by (6.3.20) and (6.3.19).

Proof It follows from Definition 4, (6.3.19) and (6.3.20) that

$$P^a = mU^a = m(\gamma Z^a + \gamma u^a) = EZ^a + p^a.$$

□

Remark 7 Equation (6.3.32) indicates that the 3-momentum p^a and the energy E are respectively the spatial and time components of the 4-momentum P^a , the latter of which can be expressed as

$$E = -P^a Z_a. \quad (6.3.33)$$

[This can be easily seen by contracting Z^a with (6.3.32)]. The concept of the 4-momentum of a point mass unifies two different concepts—the energy and momentum of a point mass—organically into one physical quantity, which is independent of the observer (P^a is absolute). However, the way of decomposing P^a into time and spatial components depends on the observer, and thus is relative. If there is no observer making a local measurement, then the 4-momentum still exists objectively, but the energy and 3-momentum are meaningless. Now we can further understand why most modern literature only uses the (rest) mass m and (total) energy E —they are two fundamentally different types of quantity. The mass m of a point mass (e.g., an electron) is an invariant (just like its electric charge), which reflects an intrinsic

property of a point mass. The energy E of a point mass depends on the observer (and thus is not an invariant). The energy measured by an instantaneous rest observer is the rest energy; although it has the same value as the mass, they are not quantities of the same type [mass is an invariant, while the rest energy is a special case of an observer dependent quantity (energy)].

Remark 8 It is easy to derive the relation of mass, energy and 3-momentum from (6.3.32) as follows:

$$P^a P_a = (EZ^a + p^a)(EZ_a + p_a) = -E^2 + p^2,$$

where p stands for the magnitude of the 3-momentum. On the other hand, $P^a P_a = mU^a mU_a = -m^2$, and therefore

$$E^2 = m^2 + p^2, \quad (6.3.34)$$

which is exactly the well-known formula $E^2 = m^2 c^4 + p^2 c^2$ when $c = 1$.

[Optional Reading 6.3.1]

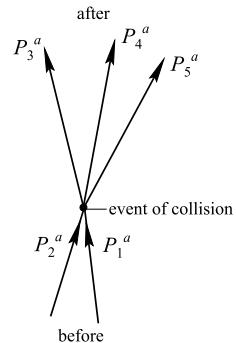
The laws of the conservation of the energy and 3-momentum in a collision process are theoretical hypotheses that have been verified by countless experiments, which can be expressed together in terms of the 4-dimensional language as: the total 4-momentum is conserved in the collision (the law of the conservation of 4-momentum). The “collision” is quite general, including all the interactions that happen at the same spacetime point; the particles involved can be either point masses or photons (see the paragraph before Optional Reading 6.6.5 for the definitions of the energy and 3-momentum of a photon), and the particle numbers before and after the collision can be different (see Fig. 6.29). Let P^a and \bar{P}^a represent the sum of the 4-momentum of all the particles before and after the collision, respectively. (For Fig. 6.29 they are $P^a = P_1^a + P_2^a$, $\bar{P}^a = P_3^a + P_4^a + P_5^a$). Then, the law of the conservation of 4-momentum can be expressed as $P^a = \bar{P}^a$. Note that this kind of vector equation is endowed with Lorentz covariance, and there is no need to worry anymore about whether the energy or 3-momentum is conserved in one frame but not conserved in another frame. The fact that the energy and 3-momentum are respectively the “time component” and “spatial component” of the 4-momentum is significant in many aspects. As an example, let us show that merely from the conservation of 3-momentum one can derive the conservation of 4-momentum, and thus the conservation of energy. First we choose an instantaneous observer (p, Z^a) such that Z^a is parallel to P^a , then P^a has no spatial component relative to Z^a , i.e., the 3-momentum $p^a = 0$. From the conservation of 3-momentum we know that $\bar{p}^a = 0$, i.e., \bar{P}^a has no spatial component relative to Z^a either, and hence \bar{P}^a and P^a at most only differ by multiplication by a numerical factor (denoted by σ): $\bar{P}^a = \sigma P^a$. Choose another instantaneous observer (p, Z'^a) (where Z'^a and P^a are not parallel), and let h'^a_b represent the projection map determined by Z'^a . Then, the 3-momentum \bar{p}'^a of \bar{P}^a relative to Z'^a satisfies $\bar{p}'^a = h'^a_b \bar{P}^b = \sigma h'^a_b P^b = \sigma p'^a$, and the fact that 3-momentum is conserved in any inertial frame (which is the key point) assures that $\bar{p}'^a = p'^a$, and hence $\sigma = 1$. Therefore, $P^a = \bar{P}^a$, i.e., the 4-momentum is conserved.

[The End of Optional Reading 6.3.1]

Definition 5 The 4-acceleration of a point mass is defined as

$$A^a := U^b \partial_b U^a, \quad (6.3.35)$$

Fig. 6.29 Two particles becomes three particles after a collision



where U^a is the 4-velocity of the point mass, and ∂_b is the derivative operator associated with η_{ab} (i.e., $\partial_a \eta_{bc} = 0$).

Remark 9 By definition we can see that ① 4-acceleration is absolute; ② $A^a = 0$ is equivalent to $U^b \partial_b U^a = 0$ (the world line being a geodesic), i.e., the point mass undergoes inertial motion. Thus, a necessary and sufficient condition for a point mass to experience an inertial motion (be a free point mass) is that its 4-acceleration is zero.

Proposition 6.3.4 *The 4-acceleration A^a at each point on the world line of a point mass is orthogonal to the 4-velocity U^a , i.e., $A^a U_a = \eta_{ab} A^a U^b = 0$.*

Proof Exercise 6.12. (Hint: use $U^b \partial_b (U^a U_a) = 2U_a U^b \partial_b U^a$). □

Unlike the 3-velocity u^a , the 3-acceleration of a point mass L cannot be determined just by one observer G , since to determine the 3-acceleration of L at p (the intersection of G and L) one needs to compare the 3-velocities of L at p and at another point p' sitting next to p on L , while the latter in general is not an intersection of G and L . This difficulty can be overcome by means of a coordinate system: one can define the 3-acceleration of L at an arbitrary p on it relative to any coordinate system (called the “coordinate 3-acceleration”). The most commonly used one should be the 3-acceleration of L relative to an inertial coordinate system.

Definition 6 Suppose the parametric equations of the world line $L(\tau)$ of a point mass in an inertial coordinate system $\{t, x^i\}$ are $t = t(\tau)$, $x^i = x^i(\tau)$, then its **3-acceleration** relative to this system is defined as

$$a^a := \frac{d^2 x^i(t)}{dt^2} \left(\frac{\partial}{\partial x^i} \right)^a, \quad (6.3.36)$$

where $x^i(t)$ is the function $x^i = x^i(t)$ obtained by combining $x^i = x^i(\tau)$ and $t = t(\tau)$ (namely the parametric equation of L with t as the parameter).

Remark 10 It is ease to see that this definition agrees with (6.3.18).

Now we discuss the relation between the 4-acceleration A^a of a point mass and its 3-acceleration a^a relative to an inertial frame \mathcal{R} .

Proposition 6.3.5 *The components of the 4-acceleration A^a in an inertial frame \mathcal{R} are*

$$A^0 = \gamma^4 \vec{u} \cdot \vec{a}, \quad A^i = \gamma^2 a^i + \gamma^4 (\vec{u} \cdot \vec{a}) u^i, \quad (6.3.37)$$

where \vec{u} and \vec{a} are respectively the 3-velocity and 3-acceleration of the point mass relative to \mathcal{R} , $\gamma \equiv (1 - u^2)^{-1/2}$, and $u \equiv (\vec{u} \cdot \vec{u})^{1/2}$.

Proof Suppose $\{(dx^\mu)_a\}$ is the dual coordinate basis of the frame \mathcal{R} , then it follows from the definition of A^a that

$$\begin{aligned} A^\mu &= A^a (dx^\mu)_a = (dx^\mu)_a U^b \partial_b U^a = U^b \partial_b [(dx^\mu)_a U^a] \\ &= U^b \partial_b U^\mu = \frac{dU^\mu}{d\tau} = \gamma \frac{dU^\mu}{dt}. \end{aligned}$$

(The third equality is because the ∂_a that satisfies $\partial_a \eta_{bc} = 0$ is exactly the ordinary derivative operator of the Lorentzian system). From (6.3.30) we can see that $U^0 = \gamma$, $U^i = \gamma u^i$, and thus

$$\begin{aligned} A^0 &= \gamma \frac{dU^0}{dt} = \gamma \frac{d\gamma}{dt}, \\ A^i &= \gamma \frac{dU^i}{dt} = \gamma \frac{d(\gamma u^i)}{dt} = \gamma^2 \frac{du^i}{dt} + u^i \gamma \frac{d\gamma}{dt} = \gamma^2 a^i + u^i \gamma \frac{d\gamma}{dt}. \end{aligned}$$

Also, from $\gamma \equiv (1 - u^2)^{-1/2}$ we get $d\gamma/dt = \gamma^3 u du/dt = \gamma^3 \vec{u} \cdot \vec{a}$. Plugging this into the two equations above yields (6.3.37). \square

Remark 11 For a free point mass we have $A^a = 0$, and from (6.3.37) we can see that its 3-acceleration relative to any inertial frame is $a^a = 0$.

Proposition 6.3.6 *The 4-acceleration of a point mass is equal to its 3-acceleration relative to an instantaneous rest inertial coordinate system.*

Proof Plugging $\vec{u} = 0$ into (6.3.37) yields $A^0 = 0$ and $A^i = a^i$. \square

Definition 7 The **4-force** on a point mass is defined as

$$F^a := U^b \partial_b P^a, \quad (6.3.38)$$

where U^a and P^a are the 4-velocity and 4-momentum of the point mass, respectively.

Equation (6.3.38) is also called (the 4-dimensional expression of) the relativistic equation of motion for a point mass, but actually it is just the definition of the 4-force.

To obtain the real physical laws, one also needs to combine (6.3.38) with the specific expression of the 4-force in each specific case.

Remark 12 In this section, we only care about the case where the (rest) mass m of the point mass remains a constant ($dm/d\tau = 0$). In this case plugging $P^a = mU^a$ into (6.3.38) we obtain $F^a = mA^a$. However, if m is changing during the motion ($dm/d\tau \neq 0$), then this conclusion does not hold, see Optional Reading 6.3.

Proposition 6.3.7 *The spatial components F^i ($i = 1, 2, 3$) of the 4-force on a point mass in an inertial coordinate system $\{x^\mu\}$ is equal to γ times the corresponding component f^i of the 3-force acting on it, and the time component F^0 of the 4-force is equal to γ times $\vec{f} \cdot \vec{u}$, the power of the 3-force. That is,*

$$F^i = \gamma f^i, \quad F^0 = \gamma \vec{f} \cdot \vec{u}, \quad (6.3.39)$$

where $\gamma \equiv (1 - u^2)^{-1/2}$, and u is the magnitude of the 3-velocity \vec{u} of the point mass with respect to this system.

Proof In $\{x^\mu\}$, the components of F^a are

$$F^\mu = F^a (dx^\mu)_a = (dx^\mu)_a U^b \partial_b P^a = U^b \partial_b [(dx^\mu)_a P^a] = U^b \partial_b P^\mu.$$

Take the i th component. It follows from (6.3.32) and (6.3.19) that

$$F^i = U^b \partial_b P^i = \frac{dp^i}{d\tau} = \frac{dp^i}{dt} \frac{dt}{d\tau} = \gamma f^i.$$

Now take the 0th component. It follows from (6.3.32) and (6.3.20) that

$$F^0 = U^b \partial_b P^0 = U^b \partial_b E = \frac{dE}{d\tau} = \frac{dE}{dt} \frac{dt}{d\tau} = \gamma \frac{dE}{dt} = \gamma \vec{f} \cdot \vec{u}.$$

□

[Optional Reading 6.3.2]

So far we only discussed the case where the (rest) mass m of the point mass is a constant ($dm/d\tau = 0$), but more generally, m may change in the motion, i.e., $dm/d\tau \neq 0$. For instance, consider a resistor in a DC circuit at rest in an inertial frame \mathcal{R} . The Joule heat (which is also a form of energy) caused by the current makes the rest energy mc^2 of the resistor to increase, and thus $dm/d\tau > 0$. In the case where $dm/d\tau \neq 0$, some previous conclusions need to be modified. Such as,

(1) Although $\vec{f} \cdot \vec{u}$ can still be called the power of the 3-force (there are also people who think it is improper to call it so), it is not equal to the rate of change of the total energy any more. The relation between them is now

$$\vec{f} \cdot \vec{u} = \frac{dE}{dt} - \frac{c^2}{\gamma} \frac{dm}{dt}. \quad (6.3.40)$$

Fig. 6.30 The triad of an observer (spatial diagram)

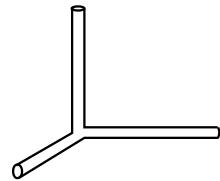
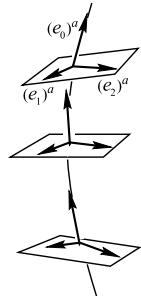


Fig. 6.31 The tetrad field along a world line



(2) The kinetic energy should be defined as the difference between the total energy γmc^2 and the rest energy mc^2 , i.e., $E_k = (\gamma - 1)mc^2$, which does not satisfy $\vec{f} \cdot \vec{u} = dE_k/dt$. In fact, if $d\mathbf{m}/d\tau \neq 0$ then $\vec{f} \cdot \vec{u}$ is equal to neither dE/dt nor dE_k/dt .

(3) The 4-force is still defined as $U^b \partial_b P^a$; however, $F^a \neq mA^a$.

(4) Proposition 6.3.7 now should be stated as

$$F^i = \gamma f^i, \quad F^0 = \gamma \frac{dE}{dt} (\neq \gamma \vec{f} \cdot \vec{u}). \quad (6.3.41)$$

[The End of Optional Reading 6.3.2]

To make a measurement, besides a standard clock, each observer also needs to be equipped with a **triad (3-dimensional frame)**. Intuitively, a triad is a frame welded by three short rods with unit length that are orthogonal to each other (see Fig. 6.30); the observer chooses which direction each rod points at, which represents a direction of the measurement. Mathematically, a triad is abstracted as three orthonormal spatial vector fields $\{(e_i)^a, i = 1, 2, 3\}$ on the observer's world line; “spatial” means that they are all orthogonal to the 4-velocity Z^a of the observer. Hence, including $(e_0)^a = Z^a$, there are four orthonormal vector fields along the observer's world line, called the **tetrad field (4-dimensional frame field)**, see Fig. 6.31. Later, unless stated otherwise, when we talk about a tetrad field it will refer to a right-handed tetrad field. Recall that a reference frame is formed by infinitely many observers filled in the spacetime, and at each spacetime point there is one and only one observer's world line passing through it. Thus, given a reference frame, we will have a tetrad field in the whole spacetime (or in an open subset of it). Any tensor at any spacetime point can be expressed in terms of the tetrad at this point as a basis.

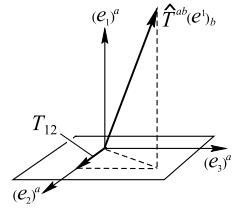
Previously, when we talked about an observer we meant a timelike curve, while this is not sufficient in many situations, in which one also needs to add a requirement on the tetrad, i.e., an observer is a timelike curve with a tetrad field defined along it. The precise definition of an inertial observer is then: an inertial observer is a non-rotating observer undergoing inertial motion. “Undergoing inertial motion” means the world line is a geodesic (before, this was the only requirement for an inertial observer), while “non-rotating” is the requirement for the tetrad field on the world line. Intuitively, suppose two boys A and B sit on two chairs on the ground. A sits on a regular chair while B sits on a swivel chair (whose base is fixed on the ground) and keeps rotating. Then, A can be viewed as an inertial observer while B cannot (since he is rotating). Note that although the concept of an observer requires us to treat the two boys as point-like (and thus each of them is represented by a world line, and both are geodesics), the question whether or not there is rotation only depends on whether or not the direction of each spatial basis vector at each point changes along the curve, and thus it is a meaningful question (more detailed discussion of this can be found in Sect. 7.3). For instance, suppose \mathcal{R} is an inertial coordinate system, treat each of the t -coordinate lines in it as the world line of an observer, and choose the inertial coordinate basis as the tetrad field along the curve. Then, intuitively speaking (or according to the precise definition in Sect. 7.3), each observer is non-rotating. Normally when we talk about an inertial observer in an inertial frame, we often assume that the inertial coordinate basis is used as the tetrad. Correspondingly, to determine an instantaneous observer, one needs not only the ingredients of p and Z^a , but also replacing Z^a by an orthonormal tetrad at p . Therefore, an instantaneous observer should be represented by $(p, (e_\mu)^a)$, where $(e_0)^a = Z^a$. When it is not necessary to emphasize the tetrad, an instantaneous observer can still be represented by (p, Z^a) .

6.4 The Energy-Momentum Tensor of Continuous Media

When discussing media that are continuously distributed (gases, liquids, solids, plasma, etc.), we care not about the behavior of any specific particle, but about the statistical average over all of the particles. We are interested in the energy/momentum density and energy/momentum flux density, etc. at each point of space rather than the energy and momentum of any individual particle. Thus, a continuous medium is similar to an electromagnetic field in many aspects, and we call it a **matter field**. Suppose m is the rest mass in a macroscopically small volume V , the content of which has a 3-velocity \vec{u} relative to an inertial frame, then its 3-momentum is $\vec{p} = \gamma m \vec{u} = (E/c^2)\vec{u}$, where E is its energy and the meaning of γ is self-evident. Dividing the whole equation by V yields

$$\text{3-momentum density} = \frac{1}{c^2} \text{energy density} \times \vec{u} = \frac{1}{c^2} \text{energy flux density}. \quad (6.4.1)$$

Fig. 6.32 T_{12} is the second component of the force from the matter below the area element on the matter above, $\hat{T}^{ab}(e^1)_b$ is the 3-momentum flux density along the $(e_1)^a$ -direction (see optional reading)



(The second equality can be understood by means of the following example: in electromagnetism, suppose ρ is the charge density and \vec{u} is the velocity of the charge carrier, then the electric current density is $\vec{j} = \rho\vec{u}$). When we take $c = 1$, the 3-momentum density is then equal to the energy flux density.

The expressions for the energy density, momentum density, energy flux density (Poynting vector) and momentum flux density can be found in textbooks on electrodynamics, where the energy flux density is equal to the momentum density times c^2 . Just like the energy and 3-momentum of a particle together form the 4-momentum vector P^a , these density quantities of an electromagnetic field form a tensor T_{ab} of type $(0, 2)$, called the **energy-momentum tensor**, which is a tensor field on 4-dimensional Minkowski spacetime, and all kinds of 3-dimensional densities are nothing but different components of T_{ab} . In fact, just like electromagnetic fields, each matter field also has their own energy-momentum tensor T_{ab} , which has the following important properties and physical meanings:

1. $T_{ab} = T_{ba}$.
2. For any matter field that is closed (without interaction with the outside) we have $\partial^a T_{ab} = 0$. We will see below that this is exactly the manifestation of conservation of the energy, 3-momentum and angular momentum (the conservation of angular momentum also requires $T_{ab} = T_{ba}$).
3. For an arbitrary instantaneous observer $(p, (e_\mu)^a)$, $(e_0)^a = Z^a$ we have
 - (a) $\mu \equiv T_{ab} Z^a Z^b = T_{00}$ is the energy density measured by this observer;
 - (b) $w_i \equiv -T_{ab} Z^a (e_i)^b = -T_{0i}$ is the i -component of the 3-momentum density (energy flux density) measured by this observer;
 - (c) $T_{ab}(e_i)^a (e_j)^b = T_{ij}$ is the ij -component of the 3-stress tensor measured by this observer. For instance, take a spatial unit area element perpendicular to $(e_1)^a$ (Fig. 6.32 is the spatial diagram), then T_{12} is equal to the second component of the force exerted from the matter below the area element on the matter above (see textbooks on the theory of elasticity).

Thus, the energy-momentum tensor T_{ab} is absolute, while the energy density, 3-momentum density, etc. are relative.

[Optional Reading 6.4.1]

Since $\{(e_i)^a\}$ is orthonormal, it is not difficult to show that $T^{ij} = T_{ij}$. Suppose $\{(e^\mu)_a\}$ is the dual basis of $\{(e_\mu)^a\}$, let us discuss the physical meaning of the spatial tensor field $\hat{T}^{ab} \equiv T^{ij}(e_i)^a (e_j)^b$ [or $\hat{T}_{ab} \equiv T_{ij}(e^i)_a (e^j)_b$]. Let ΔS represent the spatial unit area element perpendicular to $(e_i)^a$ (i is any of 1, 2, 3). It follows from the text above that

$$\begin{aligned} T^{ij} = T_{ij} &= j - \text{component of the force from the matter} \\ &\quad \text{on one side of } \Delta S, \text{ to the matter on the other side,} \end{aligned} \quad (6.4.2)$$

and thus \hat{T}^{ab} should be interpreted as the 3-stress tensor. On the other hand,

$$T^{ij} = \hat{T}^{ab}(e^i)_a(e^j)_b = [\hat{T}^{ab}(e^i)_b](e^j)_a = j - \text{component of } \hat{T}^{ab}(e^i)_b. \quad (6.4.3)$$

Combining (6.4.2) and (6.4.3) yields

$$\begin{aligned} j - \text{component of } \hat{T}^{ab}(e^i)_b &= j - \text{component of the force from the matter} \\ &\quad \text{on one side of } \Delta S \text{ to the matter on the other side,} \end{aligned}$$

and hence

$$\begin{aligned} \hat{T}^{ab}(e^i)_b &= \text{the force from the matter} \\ &\quad \text{on one side of } \Delta S \text{ to the matter on the other side.} \end{aligned}$$

Also, a force is nothing but the rate of change of the 3-momentum of the object which the force acts on, and the interaction between them is nothing but exchanging their 3-momenta. Thus,

$$\begin{aligned} \hat{T}^{ab}(e^i)_b &= \text{the 3-momentum crossing a unit area perpendicular to } (e_i)^a \\ &\quad \text{along the direction of } (e_i)^a \text{ in a unit time} \\ &= \text{the 3-momentum flux density along the direction of } (e_i)^a. \end{aligned}$$

The $(e_i)^a$ in the equation above can be the unit vector of any spatial direction, and so this equation indicates that the 3-momentum flux density along any spatial direction can be obtained by contracting \hat{T}^{ab} with the unit vector of this direction. Therefore, $\hat{T}^{ab} \equiv T^{ij}(e_i)^a(e_j)^b$ can be interpreted (called) as the **3-momentum flux density tensor**.

[The End of Optional Reading 6.4.1]

Definition 1 $W^a := -T^a{}_b Z^b$ is called the **4-momentum density** measured by the instantaneous observer (p, Z^a) .

Proposition 6.4.1 *The 4-momentum density W^a measured by the instantaneous observer $(p, (e_\mu)^a)$, $(e_0)^a = Z^a$ can be decomposed as follows:*

$$W^a = \mu Z^a + w^a, \quad (6.4.4)$$

where μ and $w^a \equiv w^i(e_i)^a$ are respectively the energy density and 3-momentum density measured by this observer; the latter of which is a spatial vector of this observer.

Proof The components of W^a in the frame $\{(e_\mu)^a\}$ are

$$\begin{aligned} W^0 &= W^a(e^0)_a = -T^a{}_b Z^b(-Z^a) = T_{ab} Z^b Z^a = \mu, \\ W^i &= W^a(e^i)_a = -T^a{}_b Z^b(e^i)_a = -T_{ab} Z^b (e^i)^a = w^i. \end{aligned}$$

Hence, $W^a = \mu(e_0)^a + w^i(e_i)^a = \mu Z^a + w^a$. □

Remark 1 Equations (6.4.4) and (6.3.32) are very similar: the left-hand side of the latter is the 4-momentum P^a , and the left-hand side of the former is the 4-momentum density W^a . Both equations are the $3 + 1$ decomposition of a 4-vector. However, one should notice a difference: the 4-momentum P^a is independent of the observer, while the 4-momentum density W^a depends on the observer (from Definition 1 one can see that W^a is a 4-vector that depends on the observer).

Proposition 6.4.2 $\partial^a T_{ab} = 0 \Rightarrow$ energy conservation.

Proof Suppose t, x, y, z are the coordinates for an inertial frame \mathcal{R} , and let $Z^a \equiv (\partial/\partial t)^a$. Then taking the derivative of $W^a \equiv -T^a{}_b Z^b$ yields

$$\partial_a W^a = \partial_a (-T^a{}_b Z^b) = -Z^b \partial^a T_{ab} - T^a{}_b \partial_a Z^b.$$

The first term on the right-hand side of the above equation vanishes (since $\partial^a T_{ab} = 0$), as does the second term [since $\partial_a Z^b = \partial_a (\partial/\partial t)^b = 0$], and hence

$$\partial_a W^a = 0. \quad (6.4.5)$$

Therefore,

$$0 = \partial_\mu W^\mu = \partial_0 W^0 + \partial_i W^i = \partial_0 \mu + \partial_i w^i = \frac{\partial \mu}{\partial t} + \vec{\nabla} \cdot \vec{w}. \quad (6.4.6)$$

Since μ and w^a are respectively the energy density and energy flux density measured by the frame \mathcal{R} , the equation above looks quite like the continuity equation $(\partial \rho/\partial t) + \vec{\nabla} \cdot \vec{j} = 0$ in electrodynamics. Following the reasoning of the conservation of the electric charge from the latter, one can deduce that (6.4.6) leads to the conservation of energy. \square

Remark 2 One can also derive the conservation of 3-dimensional momentum and angular momentum from $\partial^a T_{ab} = 0$, and thus $\partial^a T_{ab} = 0$ is also called the **conservation equation**.

[Optional Reading 6.4.2]

The conservation of energy can also be derived directly from (6.4.5) using the 4-dimensional version of Gauss's Theorem as follows: let Ω to be the 4 dimensional “cuboid” bounded by several hypersurfaces ($3d!$) in \mathbb{R}^4 (see Fig. 6.33, one dimension is suppressed in the figure), i.e., (a segment of) the world tube of the 3-dimensional rectangular box ω (shown in Fig. 6.34). It follows from Gauss's theorem and (6.4.5) that

$$0 = \int_{\partial\Omega} W^a n_a = \int_{\sigma_1} W^a n_a + \int_{\sigma_2} W^a n_a + \int_{\Delta} W^a n_a. \quad (6.4.7)$$

σ_1 and σ_2 are the “upper and lower bases” of Ω , and Δ represents all of the “sides” of Ω . Noticing the requirement on the direction of the normal vector in (5.5.7'), we can see that the normal vector of σ_1 , σ_2 and Δ (one of the side surfaces) is in the direction shown in Fig. 6.33. Thus,

$$\begin{aligned}\int_{\sigma_1} W^a n_a &= \int_{\sigma_1} (\mu Z^a + w^a) n_a = \int_{\sigma_1} \mu Z^a Z_a = - \int_{\sigma_1} \mu \\ &= -E_1 = -(\text{the energy of the } 3d \text{ box } \omega \text{ at } t_1),\end{aligned}$$

where (6.4.4) is used in the first equality. Similarly,

$$\int_{\sigma_2} W^a n_a = E_2 = (\text{the energy of } \omega \text{ at } t_2).$$

On the other hand,

$$\int_{\Delta_1} W^a n_a = - \int_{\Delta_1} T_{ab} Z^b (\partial/\partial x)^a = \int_{\Delta_1} w_1 = \int_{\Delta_1} w_1 \hat{\epsilon},$$

where $\hat{\epsilon}$ is the 3-dimensional volume element induced by the 4-dimensional volume element $\epsilon = dt \wedge dx \wedge dy \wedge dz$ on Δ_1 , i.e.,

$$\hat{\epsilon}_{abc} = (\partial/\partial x)^d (dt)_d \wedge (dx)_a \wedge (dy)_b \wedge (dz)_c = -(dt)_a \wedge (dy)_b \wedge (dz)_c.$$

Hence,

$$\int_{\Delta_1} W^a n_a = - \int_{\Delta_1} w_1 dt \wedge dy \wedge dz = \int_{\Delta_1} w_1 dt dy dz = \int_{t_1}^{t_2} \int_{y_1}^{y_2} \int_{z_1}^{z_2} (w_1 dy dz) dt, \quad (6.4.8)$$

where the minus sign is dropped in the second equality because $\{t, y, z\}$ is a left-handed coordinate system measured by $\hat{\epsilon} = -dt \wedge dy \wedge dz$. Recalling that w_1 is the energy flux density along the direction of $(\partial/\partial x)^a$, we can see that $\int_{y_1}^{y_2} \int_{z_1}^{z_2} (w_1 dy dz) dt$ is the energy flowing out of the side wall S_1 of ω within a time dt , and hence the right-hand side of (6.4.8) is the energy flowing out of ω from the side wall S_1 (see Fig. 6.34) in a time $t_2 - t_1$, and $-\int_{\Delta} W^a n_a$ is the energy flowing into ω from each side wall in $t_2 - t_1$, i.e., the energy increase in this period of time. Therefore, (6.4.7) indicates that:

$$\begin{aligned}&\text{the energy increase of the box } \omega \text{ in } t_2 - t_1 \\ &= \text{the energy flowing into } \omega \text{ from each side wall.}\end{aligned}$$

Thus, the energy is conserved.

Finally, we shall point out one subtlety in the derivation of $\int_{\sigma_1} W^a n_a = -E_1$. The expression $\int_{\sigma_1} W^a n_a$ is an abbreviation of $\int_{\sigma_1} (W^a n_a) \hat{\epsilon}$, where $\hat{\epsilon}$ is the induced volume element on σ_1 , which seems should be expressed according to (5.5.6) as $\hat{\epsilon}_{abc} = n^d \varepsilon_{dabc}$. However, the n^a in (5.5.6) is an outgoing unit normal vector, which differs from the n^a here (see Fig. 6.33) by a minus sign, and thus $\hat{\epsilon}$ should be expressed using the n^a here as

$$\begin{aligned}\hat{\epsilon}_{abc} &= -n^d \varepsilon_{dabc} = -(\partial/\partial t)^d (dt)_d \wedge (dx)_a \wedge (dy)_b \wedge (dz)_c \\ &= -(dx)_a \wedge (dy)_b \wedge (dz)_c.\end{aligned}$$

This indicates that the coordinate system $\{x, y, z\}$ on σ_1 is left-handed measured by $\hat{\epsilon}$. Therefore,

$$\int_{\sigma_1} (W^a n_a) \hat{\epsilon} = - \int_{\sigma_1} \mu \hat{\epsilon} = \int_{\sigma_1} \mu (dx)_a \wedge (dy)_b \wedge (dz)_c = - \int_{\sigma_1} \mu dx dy dz = -E_1.$$

Fig. 6.33 Ω is the world tube of the 3-dimensional box ω in Fig. 6.33 (with one dimension suppressed)

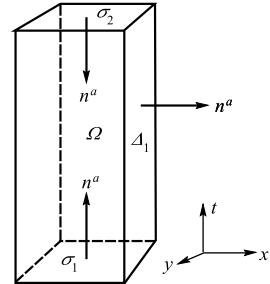
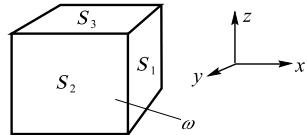


Fig. 6.34 The 3-dimensional box ω (spatial diagram)



[Since $\{x, y, z\}$ is a left-handed system, we used (5.2.6) in the third equality]. Although the conclusion is still $\int_{\sigma_1} W^a n_a = -E_1$, one should note that there are two minus signs showing up which cancel each other that assures the same result.

[The End of Optional Reading 6.4.2]

6.5 Perfect Fluid Dynamics

Definition 1. A **perfect fluid** is a matter field whose energy-momentum tensor can be expressed as

$$T_{ab} = \mu U_a U_b + p(\eta_{ab} + U_a U_b) = (\mu + p)U_a U_b + p\eta_{ab}, \quad (6.5.1)$$

where u and p are functions (scalar fields), and U^a is a future-directed timelike vector field which satisfies $U^a U_a = -1$, called the **4-velocity field** of the perfect fluid.

A fluid itself can be viewed as a reference frame. Suppose the 4-velocity $(e_0)^a$ of an instantaneous observer $(p, (e_\mu)^a)$ satisfies $(e_0)^a = U^a|_p$, then this observer is at rest relative to the fluid reference frame, and thus is called an instantaneous **rest observer**. However, to another reference frame, this observer moves with the fluid, and hence $(p, U^a|_p)$ is also called an instantaneous **comoving observer**. For a comoving observer,

$$T_{ab}(e_0)^a (e_0)^b = T_{ab} U^a U^b = (\mu + p)U_a U_b U^a U^b + p\eta_{ab} U^a U^b = (\mu + p) - p = \mu.$$

Thus, the μ in (6.5.1) is the energy density measured by a comoving observer, also called the **proper energy density**. Let $(e_i)^a$ represent the triad of a comoving observer, it follows from (6.5.1) that

$$T_{ab}(e_i)^a(e_j)^b = p\eta_{ab}(e_i)^a(e_j)^b = p\delta_{ij}.$$

Thus, the 3-dimensional stress tensor measured by a comoving observer has the matrix form

$$\begin{pmatrix} p & 0 & 0 \\ 0 & p & 0 \\ 0 & 0 & p \end{pmatrix},$$

i.e., there is only pressure but no shear stress (which is exactly an important property of a perfect fluid⁴). From $T_{11} = T_{22} = T_{33} = p$ and the arbitrariness of the triad of a comoving observer we can see that a perfect fluid is isotropic.⁵ Also, $T_{ab}(e_0)^a(e_i)^b$ indicates that the energy flux density measured by a comoving observer is zero, and thus there is no thermal conduction. All of these are important properties of a perfect fluid.

It is necessary to give an explanation of the physical meaning of the 4-velocity field U^a . A perfect fluid is a continuous medium, which is a model obtained from the statistical average over the microscopic discrete structure of the particles. Usually, a fluid volume element that is large enough microscopically while small enough macroscopically is called a **fluid particle** or **fluid point mass** [see Landau and Lifshitz (1987) p. 1; Zhou et al. (2000) pp. 15–17]. The U^a in (6.5.1) is the vector field formed by the 4-velocity of all fluid particles. A comoving observer is the observer at rest relative to a fluid particle, and a **comoving reference frame (rest reference frame)** is the reference frame of the observers whose 4-velocity field is U^a . One should note the conceptual difference between fluid particles and microscopic particles that form a fluid. This difference is especially prominent for an ideal gas (which is an example of a perfect fluid). Due to frequent collisions, the world lines of the gas molecules intersect a lot. Since the 4-velocity of a molecule has a sudden change during a collision, the world lines of the molecules are significantly distinct from Fig. 6.35, and so do not treat the U^a in (6.5.1) as the 4-velocity of a specific molecule. In fact, we have already taken the statistical average over the microscopic motion of the molecules when we regard an ideal gas as a perfect fluid, and U^a is the 4-velocity field after the average. Consider a box at rest in an inertial frame $\{t, x, y, z\}$, which contains an ideal gas in thermal equilibrium. Since there is no special direction, the average 3-velocity of the gas molecule is zero, and hence $U^a = (\partial/\partial t)^a$, whose integral curves are the t -coordinate lines as shown in 6.36. Thus, a comoving observer is not an observer moving with a gas molecule, but is the inertial observer at rest relative to the box.

The pressure p and the mass density μ of a perfect fluid have the following well-known relation:

⁴ In Newtonian hydrodynamics, a perfect fluid is defined as a fluid with no thermal conductivity or viscosity (and thus no shear stress for a rest observer); see Landau and Lifshitz (1987) p. 3.

⁵ As long as there exists a reference frame, in which the measurement of a fluid has no directional preference, then the fluid is said to be **isotropic**. We have shown that a comoving frame meets this requirement, so a perfect fluid is isotropic.

Fig. 6.35 The tangent vectors of the world lines of fluid particles form the fluid 4-velocity field U^a

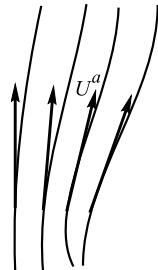
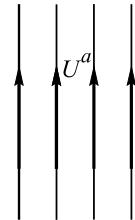


Fig. 6.36 The 4-velocity field of an ideal gas (as a perfect fluid)



$$p = \frac{\mu \bar{u}^2}{3}, \quad (6.5.2)$$

where \bar{u}^2 is the average of the square of the random motion velocity of each molecule. Since $\bar{u}^2 \ll c^2$, we have $(p/c^2) \ll \mu$, which in the unit system with $c = 1$ is $p \ll \mu$, i.e., the pressure is much less than the density. This conclusion holds for any non-relativistic fluid, like in a hurricane $p/\mu \sim 10^{-12}$, and in the Earth's core $p/\mu \sim 10^{-10}$. However, for relativistic fluids it will be quite different. The electromagnetic radiation that reaches thermal equilibrium in an isothermal box (which is called **blackbody radiation**) can be viewed as an example of an extreme relativistic perfect fluid, where the reference frame at rest relative to the box is the rest frame (comoving frame) of the fluid. The radiation inside the box is isotropic in this frame, and thus this frame is also called the isotropic reference frame of blackbody radiation. The electromagnetic radiation in the box has many similarities with an ideal gas, and can be called a **photon gas**. The relation between the pressure p and energy density μ of an photon gas also satisfies (6.5.2) (of course the derivation is different), also now $\bar{u}^2 = 1$, and hence

$$p = \frac{\mu}{3}. \quad (6.5.3)$$

The key point for why blackbody radiation can be regarded as a perfect fluid is that, relative to the isotropic reference frame, its photons have random motions in all directions that are sufficiently disordered (see Appendix D in Volume II for details). In contrast, the light rays coming from a searchlight cannot be regarded as a perfect fluid, since there does not exist a reference frame in which these light rays are isotropic.

A perfect fluid in Newtonian mechanics obeys two important laws, namely the continuity equation that describes the rate of change of the mass density μ ,

$$\frac{\partial \mu}{\partial t} + \vec{\nabla} \cdot (\mu \vec{u}) = 0 \text{ (reflects the conservation of mass),} \quad (6.5.4)$$

and the Euler equation that describes the rate of change of the 3-velocity \vec{u} (see Optional Reading 6.5 for a derivation)

$$-\vec{\nabla} p = \mu \left[\frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \vec{\nabla}) \vec{u} \right]. \quad (6.5.5)$$

Now we will introduce the generalization of these two laws in relativistic perfect fluid mechanics. Suppose a perfect fluid has no interaction with the outside, then its energy-momentum tensor satisfies $\partial^a T_{ab} = 0$. It follows from (6.5.1) that

$$0 = \partial^a T_{ab} = U_a U_b \partial^a (\mu + p) + (\mu + p)(U^a \partial_a U_b + U_b \partial_a U^a) + \partial_b p. \quad (6.5.6)$$

This is an equality of 4-vectors, which can be projected onto the spatial and time directions of a comoving observer. Contracting U^b with the equation above yields

$$0 = U^b \partial^a T_{ab} = -U_a \partial^a (\mu + p) + (\mu + p)(U^b U^a \partial_a U_b - \partial_a U^a) + U^b \partial_b p.$$

Noticing

$$U^b U^a \partial_a U_b = \frac{1}{2} U^a \partial_a (U^b U_b) = 0 \quad (\text{since } U^b U_b = -1 = \text{constant}),$$

we have

$$U^a \partial_a \mu + (\mu + p) \partial_a U^a = 0. \quad (6.5.7)$$

This is the projection of (6.5.6) in the time direction. To find the spatial projection, we contract the projection map $h_c{}^b = \delta_c{}^b + U_c U^b$ with (6.5.6) and obtain

$$(\mu + p) U^a \partial_a U_c + \partial_c p + U_c U^b \partial_b p = 0. \quad (6.5.8)$$

Equations (6.5.7) and (6.5.8) are the relativistic equations of motion for a perfect fluid. A perfect fluid with zero pressure is called a **dust**. For a dust, (6.5.8) can be simplified as $U^a \partial_a U_c = 0$, and thus the world line of a dust particle is a geodesic. This is pretty natural since $p = 0$ indicates that there is no force exerted on the particle. To find the non-relativistic approximation of (6.5.7) and (6.5.8), we choose an arbitrary inertial frame $\{t, x^i\}$ and make the 3 + 1 decomposition for U^a [see (6.3.21)]:

$$U^a = \gamma [(\partial/\partial t)^a + u^a] \cong (\partial/\partial t)^a + u^a, \quad (6.5.9)$$

where u^a is the 3-velocity of the fluid in this system, and $\gamma = -(\partial/\partial t)^a U_a$ is approximated as 1 in the non-relativistic limit. Plugging (6.5.9) into (6.5.7) and noticing that $p \ll \mu$, we get (the approximation symbol is omitted from now on)

$$0 = \left(\frac{\partial}{\partial t} \right)^a \partial_a \mu + u^a \partial_a \mu + \mu \partial_a u^a = \frac{\partial \mu}{\partial t} + \partial_a (\mu u^a).$$

Since u^a is a spatial vector in the inertial frame we are using, $\partial_a (\mu u^a) = \partial_i (\mu u^i) = \vec{\nabla} \cdot (\mu \vec{u})$, and hence the equation above is exactly the continuity equation (6.5.4). Contracting $(\partial/\partial x^i)^c$ with (6.5.8) and noticing (6.5.9) and $p \ll \mu$, we get

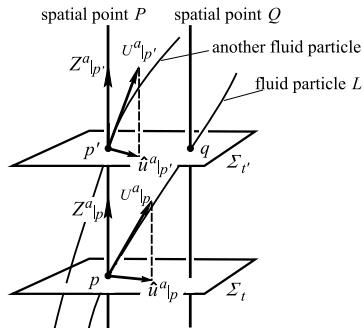
$$\begin{aligned} 0 &= \mu \left[\left(\frac{\partial}{\partial t} \right)^a \partial_a u_i + u^a \partial_a u_i \right] + \left(\frac{\partial}{\partial x^i} \right)^c \partial_c p + u_i \left[\left(\frac{\partial}{\partial t} \right)^b + u^b \right] \partial_b p \\ &= \mu \left(\frac{\partial u_i}{\partial t} + u^a \partial_a u_i \right) + \frac{\partial p}{\partial x^i} + u_i \frac{\partial p}{\partial t} + u_i u^j \frac{\partial p}{\partial x^j}. \end{aligned}$$

In the non-relativistic case ($u \ll 1$) we also have $u_i \partial p / \partial t \ll \partial p / \partial x^i$ and $u_i u^j \partial p / \partial x^j \ll \partial p / \partial x^i$, and hence the last two terms in the equation above can be neglected compared with $\partial p / \partial x^i$. Written in the form of a 3-vector equation this is exactly the Euler equation (6.5.5).

[Optional Reading 6.5.1]

The readers who have learned Newtonian fluid mechanics will know that there exists two descriptions of a fluid, namely the Lagrangian approach and the Eulerian approach [see, for example, Zhou et al. (2000)], the former of which focuses on fluid particles (the spatial trajectory of a fluid particle is called a pathline), while the latter of which focuses on spatial points (there is a flow velocity vector at each spatial point, and thus there is a flow velocity vector field \vec{u} in the 3-dimensional space, whose integral curves are called streamlines). One advantage of the Eulerian description is that a flow velocity field can be defined on space. Using the 4-dimensional language, one can acquire a deeper understanding of the difference and relationship between these two descriptions. In the 4-dimensional language, the world lines of fluid particles fill an open set O of the spacetime ($\forall p \in O$ there is a unique world line passing through p). Since the Lagrangian description focuses on fluid particles (point masses), in the 4-dimensional language it focuses on the world lines, whose tangent vectors U^a form a 4-dimensional vector field. That is, the Lagrangian description can be naturally transferred to 4-dimensional description. In contrast, the Eulerian description is intrinsically a $3 + 1$ -description, since the concept of a “spacetime point” already exists in the $3 + 1$ -language: it is nothing but a point p on a surface Σ_t of simultaneity in an inertial frame \mathcal{R} . In the 4-dimensional point of view, a spatial point is actually a world line of an inertial observer in \mathcal{R} , such as P in Fig. 6.37. $\forall p \in O$, let $\hat{u}^a|_p$ be the projection of $U^a|_p$ on the surface Σ_t of simultaneity passing through p , then we can see from (6.3.30) that $u^a|_p \equiv \gamma^{-1} \hat{u}^a|_p$ [where $\gamma \equiv -(U^a Z_a)|_p$] is the 3-velocity corresponding to the 4-velocity $U^a|_p$. Having a value of u^a at each point of the open set O , we obtain a spatial flow velocity field on O , whose dependency on the spacetime point can be expressed using the inertial coordinates t, x^i of \mathcal{R} as $u^a(t, x^i)$; at each surface Σ_t of simultaneity it gives rise to Euler's flow velocity vector field \vec{u} at a time t . As an example, we will now derive Euler's equation in order to further interpret this idea. Imagine a fluid particle as a small cube with a volume V . It is not difficult to show that the force \vec{f} acting on the particle satisfies $\vec{f}/V = -\vec{\nabla} p$, where p is the pressure at where the particle is. Suppose the mass of the fluid particle is m , then

Fig. 6.37 The definition of Euler's spatial flow velocity field u^a



$$-\vec{\nabla} p = \frac{\vec{f}}{V} = \frac{m d\vec{u}/dt}{V} = \mu \frac{d\vec{u}}{dt}, \quad (6.5.10)$$

where \vec{u} is the 3-velocity of the fluid particle. There are two reasons that \vec{u} changes with time: ① the 3-velocity \vec{u} of each spatial point can change with time (p and p' in Fig. 6.37 can have different u^a); ② a fluid particle can move from one spatial point to another spatial point (the mass point L in Fig. 6.37 moves from the spatial point P to Q), the way of its moving is described by the parametric equations $x^i = x^i(t)$ of its trajectory. Let $\vec{u}(t, x^i(t))$ represent the dependency of \vec{u} on t due to these two factors. Then, (6.5.10) can be expressed as

$$-\vec{\nabla} p = \mu \frac{d\vec{u}}{dt} = \mu \left[\frac{\partial \vec{u}}{\partial t} + \frac{\partial \vec{u}}{\partial x^i} \frac{dx^i(t)}{dt} \right] = \mu \left[\frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \vec{\nabla}) \vec{u} \right].$$

which is Euler's equation (6.5.5).

[The End of Optional Reading 6.5.1]

6.6 Electrodynamics

6.6.1 Electromagnetic Fields and 4-Current Densities

As is well-known, Maxwell's theory of electromagnetism is endowed with the Lorentz covariance. The goal for this section is to reformulate the main contents of electrodynamics using the 4-dimensional language.

There are two kinds of field involved in electrodynamics: ① the electromagnetic field; ② the matter field (a continuous fluid) formed by all of the charged particles. The latter of which is not only the source of the electromagnetic field, but also interacts with the electromagnetic field.

In the 4-dimensional language, the electromagnetic field is described by a 2-form field F_{ab} in Minkowski spacetime (called the **electromagnetic field tensor**). The electric field \vec{E} and magnetic field \vec{B} that are familiar to readers are two spatial vectors obtained by an observer measuring F_{ab} .

Definition 1 The electric field E^a and the magnetic field B^a measured by an instantaneous observer (p, Z^a) are defined by the following equations

$$E_a := F_{ab}Z^b, \quad B_a := -{}^*F_{ab}Z^b, \quad (E^a := \eta^{ab}E_b, \quad B^a := \eta^{ab}B_b.) \quad (6.6.1)$$

where ${}^*F_{ab}$ is the dual differential form of F_{ab} (see Sect. 5.6), which is also a 2-form field.

Proposition 6.6.1 E^a and B^a are spatial vector fields of the instantaneous observer $(p, (e_\mu)^a)$, $(e_0)^a = Z^a$, and

$$E_1 = F_{10}, \quad E_2 = F_{20}, \quad E_3 = F_{30}; \quad B_1 = F_{23}, \quad B_2 = F_{31}, \quad B_3 = F_{12}. \quad (6.6.2)$$

Proof Since $F_{ab} = F_{[ab]}$, $Z^a Z^b = Z^{(a} Z^{b)}$, and ${}^*F_{ab} = {}^*F_{[ab]}$, we have

$$E_a Z^a = F_{ab} Z^a Z^b = 0, \quad B_a Z^a = -{}^*F_{ab} Z^a Z^b = 0,$$

and thus E^a and B^a are spatial vectors of the instantaneous observer (p, Z^a) . Since

$$E_i = E_a (e_i)^a = F_{ab} Z^b (e_i)^a = F_{ab} (e_0)^b (e_i)^a = F_{i0},$$

we have $E_1 = F_{10}$, $E_2 = F_{20}$, $E_3 = F_{30}$. Also, since

$$B_i = B_a (e_i)^a = -{}^*F_{ab} Z^b (e_i)^a = -\frac{1}{2} \varepsilon_{abcd} F^{cd} (e_0)^b (e_i)^a = \frac{1}{2} \varepsilon_{0icd} F^{cd} = \frac{1}{2} \varepsilon_{0ijk} F^{jk},$$

we have $B_1 = \frac{1}{2} (\varepsilon_{0123} F^{23} + \varepsilon_{0132} F^{32}) = F^{23} = F_{23}$, and similarly $B_2 = F_{31}$, $B_3 = F_{12}$. \square

From Proposition 6.6.1 we can see that the matrix constituted by the components of F_{ab} in terms of the observer's tetrad $(e_\mu)^a$ is

$$(F_{\mu\nu}) = \begin{bmatrix} 0 & -E_1 & -E_2 & -E_3 \\ E_1 & 0 & B_3 & -B_2 \\ E_2 & -B_3 & 0 & B_1 \\ E_3 & B_2 & -B_1 & 0 \end{bmatrix}. \quad (6.6.3)$$

Proposition 6.6.2 Suppose two inertial frames \mathcal{R} and \mathcal{R}' are related by the Lorentz transformation

$$t = \gamma(t' + vx'), \quad x = \gamma(x' + vt'), \quad y = y', \quad z = z'. \quad (6.6.4)$$

Then, the values (\vec{E}, \vec{B}) and (\vec{E}', \vec{B}') of the same electromagnetic field F_{ab} measured by two observers in these two frames have the following relationship:

$$\begin{aligned} E'_1 &= E_1, & E'_2 &= \gamma(E_2 - vB_3), & E'_3 &= \gamma(E_3 + vB_2); \\ B'_1 &= B_1, & B'_2 &= \gamma(B_2 + vE_3), & B'_3 &= \gamma(B_3 - vE_2). \end{aligned} \quad (6.6.5)$$

Proof Exercise 6.14. □

Proposition 6.6.3 Suppose the orthonormal tetrads of two instantaneous observers $(p, (e_\mu)^a)$ and $(p, (e'_\mu)^a)$ at p have the following relation: $(e'_2)^a = (e_2)^a$, $(e'_3)^a = (e_3)^a$. Then, the values (\vec{E}, \vec{B}) and (\vec{E}', \vec{B}') of the same electromagnetic field measured by these two observers also have the relation (6.6.5), in which $\gamma \equiv -(e_0)^a (e'_0)_a$.

Proof This proposition is only about the local measurement at p and does not involve any derivative. Choose the inertial frame \mathcal{R} such that the 4-velocity of the observer whose world line passes p is $(e_0)^a$, and choose another inertial frame \mathcal{R}' such that the 4-velocity of the observer whose world line passes p is $(e'_0)^a$. Then, the relation between \mathcal{R} and \mathcal{R}' will be (6.6.4). Hence, we have (6.6.5). □

Proposition 6.6.3 indicates that (6.6.5) holds for any two instantaneous observers at any spacetime point p that satisfy $(e'_2)^a = (e_2)^a$ and $(e'_3)^a = (e_3)^a$, which clarifies the misunderstanding that “(6.6.5) only holds for an inertial frame.”

[Optional Reading 6.6.1]

Propositions 6.6.2 and 6.6.3 can also be proved using the orthonormal frame transformation (see Fig. 6.38). According to (6.3.30), the $3+1$ decomposition of the 4-velocity $U^a \equiv (e'_0)^a$ of the instantaneous observer $(p, (e'_0)^a)$ relative to the instantaneous observer $(p, (e_0)^a)$ gives

$$(e'_0)^a = \gamma(e_0)^a + \gamma u^a.$$

Since the 3-velocity u^a is in the same direction as $(e_1)^a$, and $(e_1)^a$ is normalized, we have $u^a = u(e_1)^a$, and thus the above equation becomes

$$(e'_0)^a = \gamma(e_0)^a + \gamma u(e_1)^a. \quad (6.6.6)$$

This is the expansion of $(e'_0)^a$ in the orthonormal frame $\{(e_\mu)^a\}$. Now suppose the expansion of $(e'_1)^a$ is

$$(e'_1)^a = \alpha(e_0)^a + \beta(e_1)^a \quad (\alpha, \beta \text{ to be determined}),$$

It follows from $\eta_{ab}(e'_1)^a (e'_0)^b = 0$ and $\eta_{ab}(e'_1)^a (e'_1)^b = 1$ that $\beta = \gamma$, $\alpha = \gamma u$, and hence

$$(e'_1)^a = \gamma u(e_0)^a + \gamma(e_1)^a. \quad (6.6.7)$$

Equations (6.6.6) and (6.6.7) together with $(e'_2)^a = (e_2)^a$ and $(e'_3)^a = (e_3)^a$ are the transformation relations of the two orthonormal frames $\{(e'_\mu)^a\}$ and $\{(e_\mu)^a\}$, using which it is easy to prove (6.6.5). Take E'_2 for example:

$$\begin{aligned} E'_2 &= F'_{20} = F_{ab}(e'_2)^a (e'_0)^b = F_{ab}(e_2)^a [\gamma(e_0)^b + \gamma u(e_1)^b] \\ &= \gamma(F_{20} + uF_{21}) = \gamma(E_2 - uB_3). \end{aligned}$$

[The End of Optional Reading 6.6.1]

The sources of the electromagnetic field are electric charges and electric currents. In the 4-dimensional language, the continuously distributed electric charges and currents can be viewed as a dust formed by a large amount of charged particles [see

Fig. 6.38 The relationship between two orthonormal frames

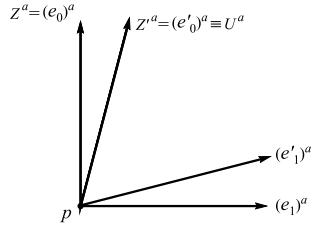
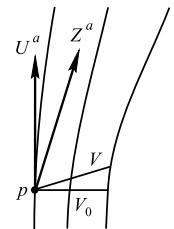


Fig. 6.39 The volumes V_0 and V measured by the comoving observer U^a and a non-comoving observer Z^a are different



Synge (1956), Chap. VIII Sect. 10, Chap. X Sect. 7]. To simplify the question, we only talk about the case where all the charged particles are of the same kind (e.g., they are all electrons), whose electric charge is e .⁶ Let U^a represents the 4-velocity field of this charged dust, then (p, U^a) is the instantaneous comoving observer at p . Suppose there are N charged particles in the small volume V_0 of the local surface of simultaneity perpendicular to U^a , then $\eta_0 = N/V_0$ is the particle number density measured by the comoving observer (called the **proper number density**). Let (p, Z^a) be an arbitrary instantaneous observer at p . This observer will regard the particle as in motion, i.e., will see a current, as long as it is not a comoving observer (as long as $Z^a \neq U^a$). Suppose the N particles above take a volume V in the local surface of simultaneity of (p, Z^a) perpendicular to Z^a (see Fig. 6.39), then from the Lorentz contraction we know that $V_0 = \gamma V$, where $\gamma \equiv -Z^a U_a$, and thus the particle number density measured by the observer (p, Z^a) is $\eta = N/V = \gamma N/V_0 = \gamma \eta_0$. Therefore, $\rho_0 \equiv e\eta_0$ and $\rho \equiv e\eta$ are, respectively, the **charge density** observed by the comoving observer (p, U^a) and that observed by an arbitrary observer (p, Z^a) , which have the relation $\rho = \gamma \rho_0$. Suppose u^a is the 3-velocity of the charged particle relative to (p, Z^a) , then $j^a := \rho u^a$ is the **3-current density** measured by (p, Z^a) . The 3-current density measured by the comoving observer is zero.

Definition 2 The **4-current density** of a stream of charged particles is defined as

$$J^a := \rho_0 U^a . \quad (6.6.8)$$

⁶ This simplification does not affect the essence of the problem. What is important is that they form a stream of particles, and unlike gas molecules which move randomly in all the directions, the value of its 4-velocity field U^a at each spacetime point is the 4-velocity of the dust particle whose world line passes through this point.

Proposition 6.6.4 J^a can be 3 + 1-decomposed by means of an instantaneous observer $(p, (e_\mu)^a)$ as follows:

$$J^a = \rho Z^a + j^a. \quad (6.6.9)$$

Proof $J^a = \rho_0 U^a = \rho_0 \gamma (Z^a + u^a) = \rho Z^a + \rho u^a = \rho Z^a + j^a.$ \square

Thus, the charge density ρ and 3-current density j^a are respectively the time component J^0 and spatial projection $h^a{}_b J^b$ of the 4-current density. The equation above can also be expressed as

$$\rho = -Z_a J^a, \quad j^i = J^i.$$

Like mass, electric charge is also a physical quantity that describes an intrinsic property of a charged particle. The charged particles and electric charges remain the same when they are not involved in any interaction. When they are interacting with other particles, the total charge must be the same before and after the interaction. This is the law of conservation of charge, which is a result confirmed by all the experiments so far. In the 3-dimensional language of electrodynamics, this law is expressed as the continuity equation: $(\partial\rho/\partial t) + \vec{\nabla} \cdot \vec{j} = 0$ (for any inertial frame). It is not difficult to see that the corresponding 4-dimensional expression is $\partial_a J^a = 0.$

6.6.2 Maxwell's Equations

In electrodynamics textbooks, the equations of motion of \vec{E} and \vec{B} are the well-known Maxwell equations. From these equations one can derive the 4-dimensional formulation of Maxwell's equations

$$\partial^a F_{ab} = -4\pi J_b, \quad (6.6.10)$$

$$\partial_{[a} F_{bc]} = 0. \quad (6.6.11)$$

In our current framework, we will treat the above two equations as the starting point, i.e., we will assume the electromagnetic field tensor obeys (6.6.10) and (6.6.11). Note that (6.6.10) already contains the law of conservation of charge, since from it we get

$$\partial^b J_b = -(4\pi)^{-1} \partial^b \partial^a F_{ab} = -(4\pi)^{-1} \partial^{(b} \partial^{a)} F_{[ab]} = 0,$$

and thus $\partial\rho/\partial t + \vec{\nabla} \cdot \vec{j} = 0$, which is exactly the conservation of charge.

Proposition 6.6.5 For any inertial frame $\{t, x, y, z\}$, from (6.6.10) and (6.6.11) one can derive the 3-dimensional formulation of Maxwell's equations

$$(a) \vec{\nabla} \cdot \vec{E} = 4\pi\rho, \quad (b) \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}, \\ (c) \vec{\nabla} \cdot \vec{B} = 0, \quad (d) \vec{\nabla} \times \vec{B} = 4\pi \vec{j} + \frac{\partial \vec{E}}{\partial t}. \quad (6.6.12)$$

The first and fourth equations here correspond to (6.6.10), and the second and third equations correspond to (6.6.11).

Remark 1 Here we adopt the geometrized Gaussian unit system (see Appendix A), in which the coefficients of the 3-dimensional Maxwell equations are slightly different from the common form.

Proof Let δ_{ab} represent the (induced) Euclidean metric on a constant- t surface of the chosen inertial frame, and let $\hat{\partial}_a$ and ∂_a represent the derivative operators associated with the metrics δ_{ab} and η_{ab} , respectively. Setting $Z^a \equiv (\partial/\partial t)^a$, and noticing that the spatial vector E^a satisfies $E_0 = 0$, we have

$$\vec{\nabla} \cdot \vec{E} = \hat{\partial}^a E_a = \frac{\partial E^i}{\partial x^i} = \partial^a E_a = \partial^a (F_{ab} Z^b) = Z^b (-4\pi J_b) = 4\pi\rho.$$

This is (6.6.12)(a). Now we prove (6.6.12)(b). Suppose $\hat{\varepsilon}_{abc}$ is the volume element associated with δ_{ab} on the constant- t surface, then from (c) of (5.6.5) we know that

$$(\vec{\nabla} \times \vec{E})_c = \hat{\varepsilon}^{ab}{}_c \hat{\partial}_a E_b, \quad (6.6.13)$$

where $\hat{\partial}_a E_b$ can be expressed as [according to (3.1.9)]

$$\hat{\partial}_a E_b = (dx^i)_a (dx^j)_b \hat{\partial}_i E_j = (dx^i)_a (dx^j)_b \partial_i E_j. \quad (6.6.14)$$

On the other hand, $E_0 = 0$ leads to

$$\partial_a E_b = (dx^\mu)_a (dx^j)_b \partial_\mu E_j = (dx^0)_a (dx^j)_b \partial_0 E_j + (dx^i)_a (dx^j)_b \partial_i E_j.$$

Comparing the projection of the above equation on the constant- t surface with (6.6.14), and noticing that the projection of $(dx^0)_a$ vanishes and the projection of $(dx^i)_a$ are themselves, we have

$$\hat{\partial}_a E_b = h_a{}^d h_b{}^e \partial_d E_e. \quad (6.6.15)$$

Since $\hat{\varepsilon}^{ab}{}_c$ is a spatial tensor, its projection is equal to itself. Plugging (6.6.15) into (6.6.13) yields

$$(\vec{\nabla} \times \vec{E})_c = \hat{\varepsilon}^{ab}{}_c h_a{}^d h_b{}^e \partial_d E_e = \hat{\varepsilon}^{de}{}_c \partial_d E_e,$$

and hence

$$(\vec{\nabla} \times \vec{E})_c = \hat{\varepsilon}^{ab}{}_c \partial_a E_b = \hat{\varepsilon}^{ab}{}_c \partial_a (F_{be} Z^e) = Z^e \hat{\varepsilon}^{ab}{}_c \partial_a F_{be} = -Z^e \hat{\varepsilon}^{ab}{}_c \partial_e F_{ab} - Z^e \hat{\varepsilon}^{ab}{}_c \partial_b F_{ea},$$

where in the last step we used (6.6.11) and the antisymmetry of F_{ab} . Also, the second term of the right-hand side of this equation is equal to $-\hat{\varepsilon}^{ab}_c \partial_a (F_{be} Z^e)$, i.e., is equal to $-(\vec{\nabla} \times \vec{E})_c$, and hence

$$2(\vec{\nabla} \times \vec{E})_c = -Z^e \hat{\varepsilon}^{ab}_c \partial_e F_{ab}. \quad (6.6.16)$$

Suppose ε_{abcd} is the volume element associated with η_{ab} , then it follows from (5.5.6) that

$$\hat{\varepsilon}_{cab} = Z^d \varepsilon_{dcab}, \quad (6.6.17)$$

and hence (6.6.16) becomes

$$2(\vec{\nabla} \times \vec{E})_c = -Z^e Z^d \varepsilon_{dc}^{ab} \partial_e F_{ab} = -Z^e \partial_e (\varepsilon_{dc}^{ab} F_{ab} Z^d) = -Z^e \partial_e (2^* F_{dc} Z^d) = -2Z^e \partial_e B_c.$$

Thus,

$$(\vec{\nabla} \times \vec{E})_i = \left(\frac{\partial}{\partial x^i} \right)^c (\vec{\nabla} \times \vec{E})_c = -Z^e \partial_e B_i = -\frac{\partial B_i}{\partial t},$$

and therefore

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}.$$

The derivation of the other two Maxwell's equations are left to the reader in Exercise 6.16. \square

Remark 2 The 4-dimensional formulation of Maxwell's equations is explicitly Lorentz covariant, and is independent of the reference frame. The 3-dimensional formulation of Maxwell's equations is also Lorentz covariant, but it is not obvious to see. Also, the 3-dimensional formulation only holds for inertial frames; for a non-inertial frame, the equations derived from (6.6.10) and (6.6.11) will be different from the regular 3-dimensional Maxwell equations.

[Optional Reading 6.6.2]

As a volume element associated with the induced metric $h_{ab} = \eta_{ab} + Z_a Z_b$ on the constant- t surface, $\hat{\varepsilon}_{cab}$ can only be determined up to a minus sign (see the end of Optional Reading 5.5.1), i.e., $-Z^d \varepsilon_{dcab}$ can also be taken as $\hat{\varepsilon}_{cab}$. Only after we take the orientation of the constant- t surface into consideration can $\hat{\varepsilon}_{cab}$ be uniquely determined as $Z^d \varepsilon_{dcab}$. Unlike the situation when we discuss Gauss's theorem, here there does not naturally exist a manifold N with boundary such that the constant- t surface can be treated as the boundary ∂N , and thus one cannot say whether its normal vector Z^a is ingoing or outgoing. Equivalently, the constant- t surface now does not have any induced orientation. The reason we write $\hat{\varepsilon}_{cab}$ as $Z^d \varepsilon_{dcab}$ rather than $-Z^d \varepsilon_{dcab}$ is based on the following consideration: the 3-dimensional formulation of Maxwell's equation $\vec{\nabla} \times \vec{E} = -\partial \vec{B}/\partial t$ involves curl, and the condition for it to hold is that the chosen Cartesian coordinate system $\{x, y, z\}$ is right-handed (otherwise we have $\vec{\nabla} \times \vec{E} = \partial \vec{B}/\partial t$), i.e., the spatial orientation needs to be compatible with $dx \wedge dy \wedge dz$. Noting that $\varepsilon_{dcab} = (dt)_d \wedge (dx)_a \wedge (dy)_b \wedge (dz)_c$ and $Z^d = (\partial/\partial t)^d$, we know that the volume element $\hat{\varepsilon}_{cab} = (dx)_a \wedge (dy)_b \wedge (dz)_c$, which is compatible with the needed orientation.

[The End of Optional Reading 6.6.2]

6.6.3 Lorentz 4-Force

As we have pointed out previously, charged particles are the sources of the electromagnetic field (manifested by J^a), whose effect on the electromagnetic field F_{ab} is reflected by (6.6.10). Conversely, there are also forces exerted from the electromagnetic field on the charged particles, namely the Lorentz force

$$\vec{f} = q(\vec{E} + \vec{u} \times \vec{B}), \quad (6.6.18)$$

where q and \vec{u} represent respectively the electric charge and 3-velocity of the point mass. Combining the above equation and the definition of the 3-force $\vec{f} = d\vec{p}/dt$ yields the equation of motion of a charged particle in an electromagnetic field (assuming no other force)

$$\frac{d\vec{p}}{dt} = q(\vec{E} + \vec{u} \times \vec{B}). \quad (6.6.19)$$

It should be pointed out that the equation above is Lorentz covariant (although it is hard to see explicitly), which is also a manifestation of the conclusion “Maxwell’s theory of electromagnetism is endowed with Lorentz covariance”. That is, for another inertial frame \mathcal{R}' , the equation of motion of the same point mass will have the same form as (6.6.19), only the quantities that depend on the reference frame need to be labeled by ', i.e.,

$$\frac{d\vec{p}'}{dt'} = q(\vec{E}' + \vec{u}' \times \vec{B}'). \quad (6.6.19')$$

Note that q does not need to be primed, since the electric charge of a point mass is an invariant.

Proposition 6.6.6 *Suppose a point mass has electric charge q , 4-velocity U^a and 4-momentum P^a , then the force from the electromagnetic field F_{ab} on it (called the Lorentz 4-Force) is*

$$F^a = q F^a_b U^b \quad (\text{where } F^a_b \equiv \eta^{ac} F_{cb}). \quad (6.6.20)$$

Thus, the 4-dimensional equation of motion for a point mass that only experiences the electromagnetic force is

$$q F^a_b U^b = U^b \partial_b P^a. \quad (6.6.21)$$

Proof Suppose p is a point on the world line L of a charged particle, and (p, Z^a) is an instantaneous observer whose orthonormal tetrad is $\{(e_\mu)^a\}$, where $(e_0)^a = Z^a$. All we have to prove is that the components F^i and F^0 of F^a in (6.6.20) with respect to this instantaneous observer satisfy

$$F^i = \gamma f^i, \quad (6.6.22)$$

$$F^0 = \gamma \vec{f} \cdot \vec{u}, \quad (6.6.23)$$

where $\gamma = -Z^a U_a$, f^i is the i th component of the Lorentz 3-force, and $\vec{u} \equiv u^a$ is the 3-velocity of the point mass relative to (p, Z^a) .

It follows from (6.6.20) that

$$F^a = \gamma q F^a{}_b (Z^b + u^b) = \gamma q (E^a + F^a{}_b u^b), \quad (6.6.24)$$

or

$$F_a = \gamma q (E_a + F_{ab} u^b).$$

Hence,

$$F_i = (e_i)^a F_a = \gamma q (E_i + F_{ij} u^j). \quad (6.6.25)$$

If we can show that

$$F_{ij} u^j = (\vec{u} \times \vec{B})_i, \quad (6.6.26)$$

then from (6.6.25) and (6.6.18) we immediately obtain $F_i = \gamma f_i$, namely (6.6.22). Now we will prove (6.6.26).

$$\begin{aligned} (\vec{u} \times \vec{B})_c &= \hat{\epsilon}_c{}^{ab} u_a B_b = \hat{\epsilon}_c{}^{ab} u_a (-{}^* F_{bd} Z^d) = \hat{\epsilon}_c{}^{ab} u_a (-\frac{1}{2} \varepsilon_{bd}{}^{ef} F_{ef} Z^d) = -\frac{1}{2} u^a \hat{\epsilon}_{cab} \varepsilon^{bdef} F_{ef} Z_d \\ &= \frac{1}{2} u^a Z^g \varepsilon_{gcab} \varepsilon^{defb} F_{ef} Z_d = \frac{1}{2} (-3!) u^a Z^g \delta^{[d}{}_g \delta^{e}{}_c \delta^{f]}{}_a Z_d F_{ef} = -3 u^a Z^g Z_{[g} F_{ca]} \\ &= -u^a Z^g (Z_g F_{ca} + Z_a F_{gc} + Z_c F_{ag}) = F_{ca} u^a - Z_{ca} u^a E_a, \end{aligned} \quad (6.6.27)$$

where in the second last equality we used $F_{ca} = -F_{ac}$, and in the last equality we used $Z^g Z_g = -1$, $F_{ag} Z^g = E_a$ and $u^a Z_a = 0$. It follows from (6.6.27) that

$$(\vec{u} \times \vec{B})_i = (e_i)^c (\vec{u} \times \vec{B})_c = F_{ij} u^j,$$

which is exactly (6.6.26). The second term on the right-hand side of (6.6.27) is necessary, otherwise the time component of the right-hand side would be nonvanishing, which contradicts the fact that $(\vec{u} \times \vec{B})_c$ on the left-hand side is spatial. Now we will prove (6.6.23).

$$\begin{aligned} F^0 &= (e^0)_a F^a = \gamma q (e^0)_a (E^a + F^a{}_b u^b) = -\gamma q (e_0)^a F_{ab} u^b = -\gamma q F_{0i} u^i \\ &= \gamma q E_i u^i = \gamma q [E_i + (\vec{u} \times \vec{B})_i] u^i = \gamma f_i u^i = \gamma \vec{f} \cdot \vec{u}, \end{aligned}$$

which is (6.6.23). In the second equality we used (6.6.24), in the third equality we used $(e^0)_a E^a = 0$ and $(e^0)^a = -(e_0)^a$, in the sixth equality we used the orthogonality between $\vec{u} \times \vec{B}$ and \vec{u} , and in the seventh equality we used (6.6.18). \square

6.6.4 The Energy-Momentum Tensor of an Electromagnetic Field

In the 3-dimensional formulation of electrodynamics, the energy density, energy flux density, momentum density and momentum flux density (i.e., the stress tensor) of an electromagnetic field are already clearly defined [see, e.g., Griffiths (2013) Sects. 8.1 and 8.2]. These 3-dimensional quantities can be unified into a 4-dimensional tensor (the energy-momentum tensor T_{ab} of an electromagnetic field) as

$$T_{ab} = \frac{1}{4\pi} (F_{ac} F_b^c - \frac{1}{4} \eta_{ab} F_{cd} F^{cd}), \quad (6.6.28)$$

where F_{ac} is the electromagnetic field tensor. Using the result in Exercise 5.9, one can also rewrite the equation above into a more symmetric form:

$$T_{ab} = \frac{1}{8\pi} (F_{ac} F_b^c + {}^*F_{ac} {}^*F_b^c), \quad (6.6.28')$$

where ${}^*F_{ac}$ is the dual form of F_{ac} and ${}^*F_b^c = \eta^{ac} {}^*F_{ba}$. It is not difficult to verify that this tensor has the properties 1 and 3 of an energy-momentum tensor described in Sect. 6.4. Especially, after choosing an arbitrary inertial frame, from (6.6.28') one can easily obtain that

$$T_{00} = \frac{1}{8\pi} (E^2 + B^2),$$

and from (6.6.28) one can easily obtain that (see Exercise 6.17)

$$w_i = -T_{i0} = \frac{1}{4\pi} (\vec{E} \times \vec{B})_i, \quad i = 1, 2, 3,$$

which are exactly the energy density and energy flux density (which also equals the momentum density) of the electromagnetic field measured by this inertial observer. However, the property 2 of an energy-momentum tensor in Sect. 6.4 (i.e., $\partial^a T_{ab} = 0$) needs to be clarified here. When $J^a = 0$ (source free), one can show that $\partial^a T_{ab} = 0$ from the 4-dimensional formulation of Maxwell's equation, i.e., a source-free electromagnetic field obeys the conservation laws of energy, momentum and angular momentum. However, if $J^a \neq 0$, then the T_{ab} in (6.6.28) does not satisfy $\partial^a T_{ab} = 0$ [Exercise 6.18(a)]. This is quite natural, since then there are interactions between the electromagnetic field and the charged particles, which involve the exchange of energy, momentum and angular momentum [Exercise 6.18(b)]. Nevertheless, the total energy-momentum tensor of the electromagnetic field and charged particles is still conserved.

6.6.5 Electromagnetic 4-Potential and Its Equation of Motion, Electromagnetic Waves

Since F_{ab} is a 2-form, one can rewrite Maxwell's equation (6.6.11) using the notion of exterior differentiation as $d\mathbf{F} = 0$, i.e., \mathbf{F} is a closed form. Since the background manifold is \mathbb{R}^4 , from Remark 1 of Sect. 5.1 we can see that \mathbf{F} is exact, i.e., there exists a 1-form field A_a on \mathbb{R}^4 such that $\mathbf{F} = dA$, or

$$F_{ab} = \partial_a A_b - \partial_b A_a .$$

Definition 3 A 1-form field A_a that satisfies $\mathbf{F} = dA$ is called a **4-potential** of the electromagnetic field F_{ab} .

If we decompose A_a into the time and spatial components using an arbitrary inertial frame $\{t, x^i\}$:

$$A_a = -\phi(dt)_a + a_a , \quad (6.6.29)$$

then it is not difficult to show that ϕ and a_a are respectively the scalar potential and the 3-vector potential of the electromagnetic field \mathbf{F} (Exercise 6.19).

When \mathbf{F} is given, the 4-potential will not be unique. Suppose A is a 4-potential of \mathbf{F} , and χ is an arbitrary C^2 function on \mathbb{R}^4 , then $\tilde{A} \equiv A + d\chi$ is also a 4-potential of \mathbf{F} since $dd\chi = 0$. This is known as the gauge freedom of the electromagnetic 4-potential. One can impose an additional condition $\partial^a A_a = 0$ called the **Lorenz⁷ gauge condition**. The A_a that satisfies this condition always exists, since suppose $\partial^a A_a \neq 0$, then one can always choose a function χ such that $\tilde{A} \equiv A + d\chi$ satisfies $\partial^a \tilde{A}_a = 0$, and to do so χ only has to satisfy $\partial^a \partial_a \chi = -\partial^a A_a$. Noticing that

$$\partial^a \partial_a \chi = \eta^{ab} \partial_b \partial_a \chi = -\frac{\partial^2 \chi}{\partial t^2} + \frac{\partial^2 \chi}{\partial x^2} + \frac{\partial^2 \chi}{\partial y^2} + \frac{\partial^2 \chi}{\partial z^2} ,$$

we can see that the nonzero solutions for $\partial^a \partial_a \chi = -\partial^a A_a$ not only exist, but also they are numerous.

Using the 4-potential we can reformulate Maxwell's equations. $\mathbf{F} = dA$ satisfies (6.6.11) automatically, and (6.6.10) can be expressed as

$$-4\pi J_b = \partial^a (\partial_a A_b - \partial_b A_a) = \partial^a \partial_a A_b - \partial_b \partial^a A_a . \quad (6.6.30)$$

In the second equality we used $\partial^a \partial_b A_a = \eta^{ac} \partial_c \partial_b A_a = \partial_c \partial_b (\eta^{ac} A_a) = \partial_c \partial_b A^c = \partial_b \partial_c A^c = \partial_b \partial^a A_a$. Therefore, an A_b under the Lorenz gauge will satisfy the following simple equation:

$$\partial^a \partial_a A_b = -4\pi J_b . \quad (6.6.31)$$

⁷ Named after the Danish physicist Ludwig Lorenz, not to be confused with H. A. Lorentz.

The equation above is equivalent to the d'Alembert equation for the scalar potential ϕ and the vector potential \vec{a} in the 3-dimensional formulation of electrodynamics. For a source-free electromagnetic field, this will become a wave equation

$$\partial^a \partial_a A_b = 0. \quad (6.6.32)$$

We want to find the wave solutions of the form of $A_b = C_b \cos \theta$ for (6.6.32), where θ is a real scalar field called the **phase**; C^b is a nonvanishing constant vector field (“constant” means $\partial_a C^b = 0$) called the **polarization vector**. Plugging these into (6.6.32) yields

$$\cos \theta (\partial^a \theta) \partial_a \theta + \sin \theta \partial^a \partial_a \theta = 0, \quad (6.6.33)$$

and thus all the $A_b = C_b \cos \theta$ that satisfy both

$$(\partial^a \theta) \partial_a \theta = 0, \quad (6.6.34)$$

$$\partial^a \partial_a \theta = 0 \quad (6.6.35)$$

are solutions to the wave equation (6.6.32). Now we will discuss this important kind of solution in detail.

Let $K^a \equiv \partial^a \theta$. We can expand K^a in terms of the dual coordinate basis of an inertial coordinate system:

$$(d\theta)_a = \partial_a \theta = K_a = K_\mu (dx^\mu)_a.$$

In the following, we only consider the simplest (which is also the most important) case where K^a is a constant vector field ($\partial_b K^a = 0$). In this case K^μ is a constant, and integrating the above equation yields

$$\theta = K_\mu x^\mu + \theta_0 \text{ (constant).} \quad (6.6.36)$$

To see the physical meaning of K^a , let us look at the $3 + 1$ decomposition of K^a in the inertial frame $\{t, x^i\}$:

$$K^a = \omega (\partial/\partial t)^a + k^a, \quad (6.6.37)$$

where k^a and $\omega \equiv K^0$ represent the spatial and time components of K^a , respectively. Now let

$$k_a \equiv \eta_{ab} k^b, \quad k_i \equiv k_a (\partial/\partial x^i)^a,$$

and set $\theta_0 = 0$. Then, (6.6.36) becomes

$$\theta = -\omega t + k_i x^i, \quad (6.6.38)$$

and hence $A_b = C_b \cos \theta$ can now be expressed as

$$A_b = C_b \cos(\omega t - k_i x^i). \quad (6.6.39)$$

This solution agrees with the familiar expression for a monochromatic plane wave, and therefore can be called a **monochromatic electromagnetic plane wave**. “Plane” means that the surface S_0 of constant phase at a given time t_0 , i.e., a wavefront, described by $\omega t_0 - k_i x^i = \varphi_0$ (constant), is a 2-dimensional plane in \mathbb{R}^3 . Since

$$\partial_a \varphi_0 = (d\varphi_0)_a = -k_i (dx_i)_a = -k_a,$$

we see that k^a is the normal vector of S_0 . Physically, k^a is called the wave 3-vector, which represents the direction of wave propagation, and ω is called the angular frequency of the wave. Therefore, K^a is called the **wave 4-vector**.

Now we will discuss K^a in the 4-dimensional language. Consider a hypersurface \mathcal{S} of constant phase in spacetime, i.e., $\mathcal{S} \equiv \{p \in \mathbb{R}^4 | \theta_p = \text{constant}\}$. We can easily see that K_a is the normal covector of \mathcal{S} (Theorem 4.4.2), and thus K^a is the normal vector of \mathcal{S} . On the other hand, (6.6.34) indicates that $K^a K_a = 0$, and hence K^a is a null vector field and \mathcal{S} is a null hypersurface. In addition, $K^a K_a = 0$ also gives

$$0 = \partial_b (K^a K_a) = 2K^a \partial_b K_a = 2K^a \partial_b \partial_a \theta = 2K^a \partial_a \partial_b \theta = 2K^a \partial_a K_b, \quad (6.6.40)$$

and thus the integral curves of K^a are null geodesics lying on \mathcal{S} . Also, from (6.6.35) we can see that $\partial^a K_a = 0$.

Suppose Σ_0 is the surface of simultaneity of $\{t, x^i\}$ at t_0 . Let $S_0 \equiv \mathcal{S} \cap \Sigma_0$ (see Fig. 6.40), then S_0 is the set of all the points in Σ_0 that have the same phase, namely a wavefront at t_0 in the 3-dimensional language. When K_μ is a constant, \mathcal{S} is a 3-dimensional plane (a null hyperplane) and S_0 is a 2-dimensional plane, and thus once again we see that (6.6.39) represents a plane wave. \mathcal{S} can be interpreted as the world sheet of a 2-dimensional wavefront, which describes the time evolution of the wavefront (the propagation of the wave). Suppose Σ_1 is the surface of simultaneity at $t_1 (> t_0)$, then after a time $t_1 - t_0$, S_0 will propagate to a new plane $S_1 \equiv \mathcal{S} \cap \Sigma_1$. The direction of the propagation is the direction orthogonal to S_0 in Σ_0 , and the speed of the propagation is exactly the speed of light (which is a consequence of the

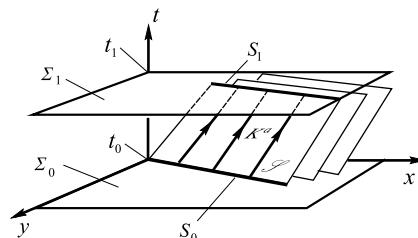


Fig. 6.40 A monochromatic electromagnetic plane wave. The world sheet of a wavefront S_0 in the 3-language is a null hypersurface \mathcal{S} . The integral curve of a normal vector K^a of \mathcal{S} represents the world line of a photon

fact that \mathcal{S} is a null hypersurface). The integral curves of the projection of K^a onto Σ_0 , i.e., the wave 3-vector k_a , are orthogonal to S_0 , which represent the direction of the wave propagation, and thus in the 3-dimensional language can be regarded as light rays. Therefore, the integral curves of K^a can be regarded as light rays in the 4-dimensional language. In this perspective, we can also naturally see that K^a deserves the name wave 4-vector.

Given a monochromatic plane wave, its wave 4-vector is also naturally given (which is a constant null vector field in \mathbb{R}^4); however, we can see from (6.6.37) that its angular frequency ω and wave 3-vector k^a will depend on the inertial frame we choose. That is, K^a is absolute, while ω and k^a are relative. Similarly, the K^a at any point p can also be decomposed in terms of an arbitrary instantaneous observer (p, Z^a) as

$$K^a = \omega Z^a + k^a, \quad (6.6.41)$$

where

$$\omega = -K^a Z_a \quad (6.6.42)$$

and k^a can be interpreted as the angular frequency and the wave 3-vector measured by this observer, respectively. From the fact that the wave 4-vector K^a is null, i.e., $K^a K_a = 0$, we can easily see the following relation between ω and k^a :

$$\omega^2 = k^a k_a = k^2. \quad (6.6.43)$$

The method of describing a monochromatic electromagnetic plane wave in terms of (either 3- or 4-dimensional) light rays is called the **geometric optics approximation**. However, the condition for a monochromatic electromagnetic plane wave is that the C_b in $A_b = C_b \cos \theta$ and $K^a \equiv \partial^a \theta$ are constant vector fields in the whole spacetime, which is an unattainable requirement and can only be a concept in theoretical models. Luckily, many electromagnetic waves in practice can be approximately be treated as this kind of wave within a certain region of spacetime, and thus can be approximated using geometric optics. Consider such an electromagnetic wave whose 4-potential can be expressed as $A_b = C_b \cos \theta$, where although C_b and $K^a \equiv \partial^a \theta$ do not satisfy $\partial_a C_b = 0$ and $\partial_a K^b = 0$ their changes with respect to the spacetime point are much “slower” than the change of the phase factor $\cos \theta$. Then, we may say that A_b is the product of the “slowly changing” amplitude C_b and the “rapidly changing” phase factor $\cos \theta$. Let \tilde{L} represent such a characteristic length that the change of C_b or K^a can only be observed when the spacetime scale is at least on the same order as \tilde{L} . Then, in a spacetime region U whose scale is smaller than \tilde{L} but $\cos \theta$ has changed by many periods in it, we can deal with this kind of electromagnetic wave using the geometric optics approximation [in the 3-dimensional language, this is to say that the spatial scale is much larger than the wavelength $\lambda \equiv 2\pi/\omega$ (where $\omega \equiv -Z^a K_a$)]. An electromagnetic wave satisfying this condition is called a **locally monochromatic plane wave**. The idea of geometric optics is to describe the propagation of an electromagnetic wave using light rays. This idea exhibits the particle nature of light, which encourages us to describe the propagation of an electromagnetic wave using

the terminology of photons. Properly, a photon is a quanta of the electromagnetic field in the theory of quantum electrodynamics (QED); classical electrodynamics can be viewed as the limit of quantum electrodynamics when the Planck constant $\hbar \rightarrow 0$.⁸ Based on this description, a locally monochromatic electromagnetic plane wave can be considered as a stream of photons, whose K^a and C^a are almost all the same. A photon can be imagined as a particle similar to a regular point mass, except that its mass $m = 0$. Now that the 4-momentum of a point mass defined by (6.3.32) no longer applies, we can instead use the corresponding wave 4-vector of the electromagnetic wave to define the 4-momentum of a photon as follows:

$$P^a := \hbar K^a \quad (\text{where } \hbar \equiv h/2\pi), \quad (6.6.44)$$

and stipulate that the world line of the photon is a null geodesic such that its affine parameter β satisfies

$$P^a = (\partial/\partial\beta)^a. \quad (6.6.45)$$

Therefore, the world lines of the photons coincide with the integral curves of the wave 4-vector of the corresponding electromagnetic wave. In terms of the $3+1$ decomposition, we can follow that of a massive particle and define the time and spatial components of a photon's 4-momentum as the energy E and the 3-momentum p^a of the photon, respectively, i.e.,

$$P^a = EZ^a + p^a. \quad (6.6.46)$$

Noticing (6.6.44), we can compare the above equation with (6.6.41) and obtain

$$E = \hbar\omega, \quad p^a = \hbar k^a, \quad (6.6.47)$$

i.e., the energy E and the 3-momentum p^a of a photon are respectively proportional to the angular frequency ω and the wave 3-vector k^a of the corresponding electromagnetic wave, with a coefficient \hbar . From $P^a P_a = 0$ one can easily see that the energy E and the magnitude p of the 3-momentum p^a has the following simple relation:

$$E^2 = p^a p_a = p^2. \quad (6.6.48)$$

[Optional Reading 6.6.3]

Equation (6.6.39) indicates that the 4-potential A_b propagates in the manner of a monochromatic plane wave, from which it is not difficult to show that the electric field \vec{E} and the magnetic field \vec{B} corresponding to an inertial frame \mathcal{R} also propagate as monochromatic plane waves, and from which we can also find some important properties of the \vec{E} wave

⁸ Note that a “photon” in the geometric optics approximation is still a classical concept since there is no procedure of quantization. A key difference between a QED photon and this classical limit is that the QED photon is not localizable, whereas the classical counterpart follows a specific ray path.

and \vec{B} wave. To proceed, we plug $A_b = C_b \cos \theta$ into $F_{ab} = \partial_a A_b - \partial_b A_a$. Noticing that $K_a \equiv \partial_a \theta$, we have

$$F_{ab} = (C_a K_b - C_b K_a) \sin \theta = 2C_a K_b \sin \theta. \quad (6.6.49)$$

Using the gauge freedom of A_b , we can simplify the computation of finding E_a and B_a from F_{ab} . Choose the Lorenz gauge condition $\partial^b A_b = 0$. Combining this with $A_b = C_b \cos \theta$ and $K^a \equiv \partial^a \theta$, we get $K^a C_a \sin \theta = 0$, and thus

$$K^a C_a = 0. \quad (6.6.50)$$

This is in fact an equivalent formulation for A_a satisfying the Lorenz gauge condition. Now let

$$C'_a = C_a + \alpha K_a \quad (\alpha = \text{constant}), \quad (6.6.51)$$

then from $K^a K_a = 0$ and (6.6.49) we can easily see that the electromagnetic field F'_{ab} corresponding to C'_a satisfies $F'_{ab} = F_{ab}$, and thus (6.6.51) is just a gauge transformation. [It follows from $K^a C_a = 0$ that (6.6.51) guarantees $K^a C'_a = 0$, and so it is also a gauge transformation within the Lorenz gauge condition]. Using the fact that the time component K^0 of K^a is nonvanishing, we can choose $\alpha = -C_0/K_0$ so that $C'_0 = 0$. Thus, one can always choose a proper gauge and render the polarization vector C^a a spatial vector. Later on we will assume the fact that C^a is a spatial vector.

Let $Z^a = (\partial/\partial t)^a$ represent the zeroth coordinate basis vector of an inertial frame, then from $E_a = F_{ab} Z^b$ and $B_a = -{}^*F_{ab} Z^b$ we can derive from (6.6.49) that

$$E_a = Z^b (C_a K_b - C_b K_a) \sin \theta = -\omega C_a \sin \theta$$

[where in the second equality we used the facts that C^a is spatial ($Z^b C_b = 0$) and $\omega = -Z^b K_b$] and also

$$B_a = -{}^*F_{ab} Z^b = -\frac{1}{2} Z^b \epsilon_{abcd} F^{cd} = \frac{1}{2} \hat{\epsilon}_{acd} 2C^{[c} K^{d]} \sin \theta = \hat{\epsilon}_{acd} C^c K^d \sin \theta,$$

where $\hat{\epsilon}$ is the volume element associated with the spatial Euclidean metric. The above two equations can be expressed in terms of “arrows” as

$$\vec{E} = -\omega \vec{C} \sin \theta = \omega \vec{C} \sin(\omega t - k_i x^i), \quad (6.6.52)$$

$$\vec{B} = \vec{C} \times \vec{k} \sin \theta = \vec{k} \times \vec{C} \sin(\omega t - k_i x^i). \quad (6.6.53)$$

Therefore,

$$\vec{B} = \hat{k} \times \vec{E}, \quad (6.6.54)$$

where \hat{k} stands for the unit vector in the direction of \vec{k} . This is exactly the often seen relation of the electric field \vec{E} , magnetic field \vec{B} and the direction \hat{k} of propagation.

Since $C_0 = 0$, the condition $K^a C_a = 0$ can now be rewritten as $k^a C_a = 0$, and thus the 3-vector \vec{C} is perpendicular to \vec{k} . And from (6.6.52) we also know that \vec{E} is parallel to \vec{C} , and hence the electric field \vec{E} is perpendicular to the direction \hat{k} , i.e., the \vec{E} wave is transverse. On the other hand, from (6.6.54) we can see that \vec{B} is perpendicular to both \hat{k} and \vec{E} , and hence the \vec{B} wave is also transverse. Conclusion: in a monochromatic electromagnetic plane wave, both the \vec{E} wave and \vec{B} wave are transverse waves with the same frequency and phase, and the vectors \vec{E} , \vec{B} and \hat{k} have the simple relation (6.6.54).

Since \vec{C} and \vec{k} are both constant vector fields, (6.6.52) and (6.6.53) represent linearly polarized light. To discuss the other polarizations, it is convenient to adopt the complex representation. First, we rewrite $A_b = C_b \cos \theta$ as

$$A_b = \operatorname{Re}(C_b e^{i\theta}) \quad (\operatorname{Re} \text{ stands for “take the real part”}), \quad (6.6.55)$$

and then generalize C^a to a constant complex vector field. This method will provide to us even richer physics. The previous proof of $K^a C_a = 0$ and the argument that C^a can be chosen to be a spatial vector field are still valid when C^a is complex, and thus the discussions and conclusions based on them still hold (including the transverse property of \vec{E} and \vec{B}). The key consequence of C^a being complex is that the linearly polarized light is generalized to elliptically polarized light. Here we will only discuss the electric field \vec{E} as an example. Now (6.6.52) should be expressed as

$$\vec{E} = \operatorname{Re}[i\omega \vec{C} e^{-i(\omega t - k_i x^i)}], \quad (6.6.52')$$

where \vec{C} is now a complex vector. Let

$$\vec{\varepsilon} \equiv i\omega \vec{C} e^{ik_i x^i}, \quad (6.6.56)$$

then (6.6.52') becomes

$$\vec{E} = \operatorname{Re}(\vec{\varepsilon} e^{-i\omega t}). \quad (6.6.57)$$

For an arbitrary observer G_0 in the frame \mathcal{R} , $\vec{\varepsilon}$ is a fixed vector while \vec{E} changes with time according to (6.6.57). Thus, the end point of the vector (arrow) \vec{E} draws a closed plane curve; we will show that it is an ellipse. Express the complex vector field $\vec{\varepsilon}$ as the sum of its real and imaginary parts:

$$\vec{\varepsilon} = \vec{\mu} + i\vec{v} \quad (\text{where } \vec{\mu} \text{ and } \vec{v} \text{ are real vector fields}). \quad (6.6.58)$$

Let β be an arbitrary real scalar field, and define real vector fields

$$\vec{m} \equiv \vec{\mu} \cos \beta + \vec{v} \sin \beta \quad \text{and} \quad \vec{n} \equiv -\vec{\mu} \sin \beta + \vec{v} \cos \beta, \quad (6.6.59)$$

then

$$\vec{\varepsilon} = \vec{\mu} + i\vec{v} = (\vec{m} + i\vec{n}) e^{i\beta}. \quad (6.6.60)$$

The advantage of introducing β is that we can choose its value such that \vec{m} and \vec{n} are orthogonal to each other, and to do so β just needs to satisfy

$$\tan 2\beta = \frac{2\vec{\mu} \cdot \vec{v}}{\mu^2 - v^2}, \quad (6.6.61)$$

where $\mu^2 \equiv \vec{\mu} \cdot \vec{\mu}$, $v^2 \equiv \vec{v} \cdot \vec{v}$, and we have supposed $\mu^2 \geq v^2$ (without loss of generality). Plugging (6.6.60) into (6.6.57) yields

$$\vec{E} = \vec{m} \cos(\omega t - \beta) + \vec{n} \sin(\omega t - \beta). \quad (6.6.62)$$

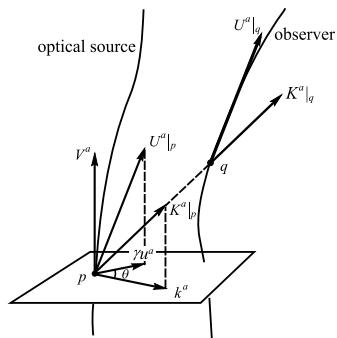
Using the orthogonality between \vec{m} and \vec{n} we can choose an inertial coordinate system $\{t, x^i\}$ in the inertial reference frame \mathcal{R} according to the following two requirements: ① Let G_0 be the origin of the spatial coordinates; ② the x - and y -axes point to the directions of \vec{m} and \vec{n} , respectively. Then, the three coordinate components of \vec{E} are accordingly,

$$E_1 = m \cos(\omega t - \beta), \quad E_2 = n \sin(\omega t - \beta), \quad E_3 = 0, \quad (6.6.63)$$

where $m \equiv (\vec{m} \cdot \vec{m})^{1/2}$, $n \equiv (\vec{n} \cdot \vec{n})^{1/2}$. From these we can easily find that

$$\frac{E_1^2}{m^2} + \frac{E_2^2}{n^2} = 1. \quad (6.6.64)$$

Fig. 6.41 The figure for discussing the Doppler effect on a light wave



Thus, as time goes on, the end point of the vector \vec{E} will draw an ellipse in the xy -plane, and therefore \vec{E} indeed represents elliptically polarized light. When $m = n$, it becomes circularly polarized light, and when m or n is zero, it goes back to linearly polarized light.

[The End of Optional Reading 6.6.3]

6.6.6 The Doppler Effect on a Light Wave

With the knowledge above (especially the $3 + 1$ decomposition of the 4-velocity U^a and the wave 4-vector K^a), the discussion of the Doppler effect on a light wave in special relativity now becomes very accessible.

Suppose an observer and a light source are undergoing arbitrary motions (their world lines are arbitrary timelike curves), and their 4-velocities are U^a and V^a (see Fig. 6.41), respectively. The light emitted at p by the light source is received at q by the observer. Assume that this light is a locally monochromatic plane wave (apply the geometric optics approximation). Suppose the wave 4-vector of the photon is K^a , then from (6.6.42) we can see that the angular frequency measured by V^a when emitting the light is $\omega = (-K^a V_a)|_p$, and the angular frequency measured by U^a when receiving the light is $\omega' = (-K^a U_a)|_p$. Now let us find the relation between ω and ω' .

Since in flat spacetime we have the notion of absolute parallel transport, we can parallelly transport $U^a|_q$ and $K^a|_q$ to p , and from the fact that parallel transport preserves the inner product we obtain $\omega' = (-K^a U_a)|_p$. Later on we will drop the subscript p , but one should remember that the calculation is at the point p . It follows from (6.6.41) that

$$K^a = \omega V^a + k^a.$$

Let

$$\gamma \equiv -V^a U_a,$$

then

$$U^a = \gamma V^a + \gamma u^a,$$

where γu^a is the projection of U^a onto the “spatial small plane” of (p, V^a) . Hence,

$$\omega' = -(\omega V^a + k^a)(\gamma V_a + \gamma u_a) = \gamma(\omega - k^a u_a).$$

Suppose the angle between the spatial vectors k^a and u^a is θ . It follows from (6.6.43) that

$$\omega' = \gamma \omega(1 - u \cos \theta). \quad (6.6.65)$$

This is the quantitative relation of the Doppler effect. If $\theta = 0$, i.e., the observer moves away from the light source, then (6.6.65) gives

$$\omega' = \gamma \omega(1 - u) = \sqrt{\frac{1 - u}{1 + u}} \omega < \omega, \quad (6.6.66a)$$

which represents a redshift; if $\theta = \pi$, i.e., the observer moves towards the light source, then

$$\omega' = \gamma \omega(1 + u) = \sqrt{\frac{1 + u}{1 - u}} \omega > \omega, \quad (6.6.66b)$$

which represents a blueshift; if $\theta = \pi/2$, i.e., the observer moves transversely, then the relation of the frequencies is

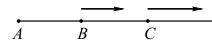
$$\omega' = \gamma \omega, \quad (6.6.66c)$$

which is called the transverse Doppler effect. The above are all the Doppler effects for a rest light source, following which one can also discuss the Doppler effects for a rest observer (Exercise 6.20).

Exercises

- ~6.1. The relative speed between two inertial observers is $u = 0.6c$. Both of their clocks C and C' are zeroed when they meet each other. Use a spacetime diagram to discuss the following questions: (a) In the inertial reference frame of C (according to its judgement of simultaneity), what is the reading of C' when the reading of C is 5 μs ? (b) When the reading of C is 5 μs , what is the actual reading of C' seen by the observer carrying C ?
- ~6.2. A celestial object is moving away from us with a constant speed $0.8c$ straight forward. The light flash it radiates has a period of 5 days when detected by us. Using a spacetime diagram, find the period of the light flash measured by an observer on that celestial object.

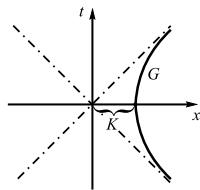
Fig. 6.42 Figure for Exercise 6.4



- ~6.3. Denote the arc length of the segments oa and oe in Fig. 6.20 as τ and τ' , respectively. (a) Express τ'/τ in terms of the relative speed of the two clocks. (b) Find the value of τ'/τ in the cases where $u = 0.6c$ and $u = 0.8c$.
- 6.4. Three inertial point masses A , B and C are aligned and moving along a straight line (see Fig. 6.42) with relative speeds $u_{BA} = 0.6c$ and $u_{CA} = 0.8c$. Suppose B thinks (measures) that C moves 60 m. Make a spacetime diagram and find the time of this process measured by A .
- ~6.5. A and B are two inertial observers in the same inertial frame that are emitting neutrons toward each other. Each neutron leaves its neutron source at a relative speed of $0.6c$. Suppose the emission rate of the source B measured by B is 10^4 s^{-1} (i.e., 10^4 per second). Using a spacetime diagram, find the emission rate of the source B measured in the reference frame of a neutron emitted by A (according to the neutron's standard clock).
- ~6.6. The mean lifetime of rest muons is $\tau_0 = 2 \times 10^{-6} \text{ s}$. A muon produced by cosmic rays is traveling down with a constant speed $0.995c$ relative to the Earth. Using a spacetime diagram, find (a) the mean lifetime of the muon measured by an Earth observer; (b) the distance that muon travels within its lifetime measured by an Earth observer.
- 6.7. From the perspective of an inertial frame \mathcal{R} , two standard clocks C_1 and C_2 at a place A start to move together with a constant speed $v = 0.6c$ after being zeroed. Both of the clocks arrive at another place B when their reading is 1 s. C_1 turns back to A with a constant speed v right after it arrives at B , while C_2 stays at B for 1 s (according to its reading) and then gets back to A with a constant speed v . There is another clock C_3 staying at A all the time, which is also zeroed at the time when C_1 and C_2 leave A . (a) Sketch the world line of C_1 , C_2 and C_3 . (b) Find the readings τ_1 , τ_2 and τ_3 of these three clocks when C_2 gets back to A .
- ~6.8. (Multiple choice). A pair of twins A and B stand still at the same spatial point in an inertial frame \mathcal{R} . At some moment when A and B are the same age, A starts to move eastward under an inertial motion with a speed u relative to the frame \mathcal{R} . A while later, B also moves eastward and catches up A with a speed $v > u$. When they meet each other again, A will be
 (1) older than B , (2) younger than B , (3) the same age as B .
- ~6.9. Two standard clocks A and B stand still at the same spatial point in an inertial frame. At some moment, A starts to move in a straight line with a speed $u = 0.6c$. 2 s later (according to the clock A), A turns around and moves back with a speed $u = 0.6c$. Both of the clocks are zeroed when they are separated. (1) Find the readings of both clocks when they meet again. (2) What is the reading of B viewed by A when A 's reading is 3 s.

- ~6.10. The equatorial speed of the Earth's rotation is about 1600 km/h. *A* and *B* are twins standing on the equator. *A* flies eastward by plane along the equator for one lap in a speed of 1600 km/h and meets *B* again when he gets back. (Ignore the effects of the gravitational fields of the Earth and the Sun. We will see in Chap. 7 that the existence of gravitational fields corresponds to a curved spacetime). (a) Sketch the world sheet of the Earth's surface and the world lines of *A* and *B* (note that the motion of *A* cancels the Earth's rotation, and thus *A* is the inertial observer). (b) Which one of *A* and *B* is younger? (c) What is their age difference? (Answer: about 10^{-7} s). NB: This experiment has been done in 1971 using cesium atomic clocks, not humans, of course. See Hafele and Keating (1972a; 1972b).
- ~6.11. A car whose rest length is $l = 5$ m moves into a garage with a constant speed $u = 0.6c$. The garage has a solid back wall. To simplify the problem, we assume the information of the car's front hitting the wall propagates in the speed of light, and each part of the car will stop once receiving this information. (a) Suppose the doorman of the garage measures that the reading of a clock *C* at the back of the car is zero, find the reading of *C* when the back of the car "learns" that the front hits the wall. (b) Find the rest length \hat{l} of the car after it comes to a complete stop. (c) Express the ratio \hat{l}/l in terms of u .
- 6.12. Prove Proposition 6.3.4.
- ~6.13. Suppose the world line of an observer is a hyperbola *G* in the tx -plane (see Fig. 6.43), which satisfies $x > 0$ and $x^2 - t^2 = K^2$ (K is a constant). Find $A^a A_a$, i.e., the magnitude square of the observer's 4-acceleration A^a . (The result is a constant, and thus *G* is called an observer undergoing constant acceleration motion. Note that the acceleration here refers to the 4-acceleration).
- ~6.14. Prove Proposition 6.6.2.
- *6.15. Suppose the electric field and the magnetic field measured from F_{ab} by an instantaneous observer are respectively E^a and B^a (also denoted by \vec{E} and \vec{B}). Show that:
- (1) $F_{ab} F^{ab} = 2(B^2 - E^2)$,
 - (2) $F_{ab}{}^* F^{ab} = 4\vec{E} \cdot \vec{B}$. Hint: one may write $F_{ab}{}^* F^{ab}$ as the expression for the components in terms of an inertial coordinate system.
- NB: this problem indicates that, although \vec{E} and \vec{B} are observer-dependent, $B^2 - E^2$ and $\vec{E} \cdot \vec{B}$ are independent of the observer. In fact, these are the only two independent invariants one can construct from F_{ab} .
- ~6.16. Prove Proposition 6.6.5 (one only needs to prove the last two Maxwell's equations).
- ~6.17. Show that the energy density and the 3-momentum density of an electromagnetic field measured by an instantaneous observer are respectively $T_{00} =$

Fig. 6.43 Figure for Exercise 6.13



$(E^2 + B^2)/8\pi$ and $w_i = -T_{i0} = (\vec{E} \times \vec{B})_i/4\pi$, $i = 1, 2, 3$. Hint: using the symmetric expression (6.6.28') for T_{ab} , one can simplify the calculation of T_{00} .

- 6.18 (a) Show that the energy-momentum tensor T_{ab} for an electromagnetic field F_{ab} whose 4-current density is J^a satisfies $\partial^a T_{ab} = -F_{bc} J^c$. (Thus, we can see that $\partial^a T_{ab} = 0$ when $J^a = 0$).
 *(b) Show that the time component of the above equation in an inertial coordinate system reflects the conservation of energy [cf. (6.108) of Jackson (1998)]; the spatial components reflects the conservation of 3-momentum [cf. (6.121) of Jackson (1998)]. Hint: rewrite $F_{bc} J^c$ as the Lorentz force density using the expression (6.6.20) for the Lorentz 4-force.
- 6.19 Show that the a^a and ϕ in (6.6.29) satisfy $\vec{B} = \vec{\nabla} \times \vec{a}$ and $\vec{E} = -\vec{\nabla} \phi - \partial \vec{a} / \partial t$, and thus are indeed the 3-vector potential and the scalar potential in electrodynamics.
- 6.20 Discuss the Doppler effects for a rest observer by following the discussion in Sect. 6.6.6. You will find that the frequency relation for the transverse Doppler effect is $\omega' = \gamma^{-1} \omega$.
- 6.21 Read Optional Reading 6.1.6. (a) Show that $\nabla_a (dt)_b = 0$, where t is the absolute time, and ∇_a is the derivative operator of the Newtonian spacetime. Hint: start from (5.7.2). (b) Suppose w^a is a spatial vector (i.e., a vector tangent to a surface of absolute simultaneity), and v^a is an arbitrary 4-vector. Show that $v^a \nabla_a w^b$ is still a spatial vector. Hint: notice that $\nabla_a t$ is the normal covector of a surface of absolute simultaneity.

References

- Griffiths, D. J. (2013), *Introduction to Electrodynamics*, Pearson, London.
 Guo, S.-H. (2008), *Electrodynamics (in Chinese)*, Higher Education Press, Beijing.
 Hafele, J. C. and Keating, R. E. (1972a), ‘Around-the-world atomic clocks: observed relativistic time gains’, *Science* **177**(4044), 168–170.
 Hafele, J. C. and Keating, R. E. (1972b), ‘Around-the-world atomic clocks: predicted relativistic time gains’, *Science* **177**(4044), 166–167.
 Jackson, J. D. (1998), *Classical Electrodynamics*, John Wiley & Sons, Inc., New York.

- Landau, L. D. and Lifshitz, E. M. (1987), *Fluid Mechanics*, Pergamon Press, Oxford.
- Misner, C., Thorne, K. and Wheeler, J. (1973), *Gravitation* W H Freeman and Company, San Francisco.
- Rindler, W. (1982), *Introduction to Special Relativity*, Clarendon Press, Oxford.
- Sachs, R. K. and Wu, H. (1977), *General Relativity for Mathematicians*, Springer-Verlag, New York.
- Synge, J. L. (1956), *Relativity: The Special Theory* North-Holland Publishing Company, Amsterdam.
- Wald, R. M. (1977), *Space, Time, and Gravity: The Theory of the Big Bang and Black Holes*, The University of Chicago Press, Chicago.
- Zhou, G., Yan, Z., Xu, S. and Zhang, K. (2000), *Fluid Mechanics (in Chinese)*, Vol. 1, Higher Education Press, Beijing.

Chapter 7

Foundations of General Relativity



7.1 Gravity and Spacetime Geometry

The principle of relativity requires that the laws of physics have the same mathematical expression in all inertial coordinate systems. When applied to special relativity, this “law of laws” requires that the mathematical expressions for the laws of physics be Lorentz covariant. Therefore, when formulating physics in the framework of special relativity, all the known laws of physics should be inspected; those that satisfy this requirement remain laws, while those that do not must be reformed until they meet this criterion. First, we inspect Maxwell’s theory of electromagnetism. Maxwell’s equations are endowed with Lorentz covariance (which can be seen more explicitly in its 4-dimensional formulation, see Sect. 6.6), and thus can be integrated into the framework of special relativity without being reformed. This is in fact not strange at all, since one of the important reasons special relativity came about is that Maxwell’s theory contradicts the notion of pre-relativity spacetime. Next, we will inspect Newton’s laws of motion. As an example, consider the law of conservation of momentum. As we pointed out at the beginning of Sect. 6.3, if the definition of momentum $\vec{p} = m\vec{u}$ is still used, then conservation of momentum violates Lorentz covariance and must be modified. By redefining momentum as $\vec{p} = m\vec{u}(1 - u^2)^{-1/2}$, the law of conservation of momentum is now Lorentz covariant, making it a valid law in the framework of special relativity. Thirdly, let us inspect Newton’s theory of universal gravity. The basic equation in Newton’s theory of gravity is Poisson’s equation $\nabla^2\phi = 4\pi\rho$, which indicates the relation between the gravitational potential ϕ and the mass density ρ .¹ This equation has Galilean covariance but not Lorentz covariance, and hence should be modified. From another perspective, Poisson’s equation $\nabla^2\phi = 4\pi\rho$ has a solution of the following form:

¹ In Chap. 6, we used ρ and μ to represent the charge density and mass density, respectively. From this chapter on, since the charge density will show up less frequently, we will follow the convention of the majority and use ρ to represent the mass density.

$$\phi(\vec{r}, t) = \iiint \frac{\rho(\vec{r}', t)}{|\vec{r}' - \vec{r}|} dV',$$

which indicates that the gravitational potential ϕ at a point \vec{r} and a time t is determined by the mass density ρ at all spatial points at t . This means that the gravitational field has an infinite speed of propagation, which obviously contradicts special relativity. Thus, Newton's theory of gravity must be modified.

The form of Newton's law of universal gravity is quite similar to Coulomb's law of electrostatics. Since James Clerk Maxwell reformulated and generalized electrostatics to such a beautiful theory of electromagnetism, it does not seem that it should be difficult to reformulate Newton's theory of gravity into a theory that fits in the framework of special relativity. However, the situation is much more complicated than this. The key point is, although the law of universal gravity and Coulomb's law are similar, there exists a "sign difference". There are two types of electric charge (positive and negative; like charges repel, while opposite charges attract); however, masses can only be positive, and hence can only attract other masses. Following the theory of electromagnetism, one might construct a gravitational theory within the framework of special relativity, and according to this theory there will be gravitational waves similar to electromagnetic waves when the gravitational field changes, which also propagate at the speed of light. Unfortunately, due to the sign difference we just mentioned, the energy carried away by such a gravitational wave has to be negative. This means that the energy of a system will increase when radiating gravitational waves, which will result in the intensity of the radiation increasing, bringing more energy into the system. This cycle inevitably leads to physically absurd consequences. Although this difficulty can be overcome by modifying the theory, new difficulties will show up. In fact, there exists far from one gravitational theory in the framework of special relativity; however, each theory has its own problems. Although one cannot completely rule out the possibility of building a satisfying gravitational theory in the framework of special relativity, Albert Einstein struck out on his own and successfully created a revolutionary gravitational theory independent of special relativity, this brand new theory is named general relativity. Interestingly, after having tried to modify a gravitational theory in the framework of special relativity in order to overcome its difficulties, what people obtained at last is a theory exactly the same as Einstein's general relativity!

There are two important factors that motivated Einstein to set up general relativity: the "universality" of gravity and Mach's principle. Here we will only introduce the former one. The meaning of the "universality" of Newtonian gravity is twofold: ① Every massive object exerts forces on other massive objects as a source of the gravitational field, and any massive object in a gravitational field will in return experience a gravitational force. (A neutral object in an electrostatic field neither exerts nor experiences any electric force, and hence the electric force is not universal). ② Any two objects with the same initial position and velocity experiencing only a gravitational force must have the same position and velocity as one another at any given moment, regardless of their mass and composition. This conclusion has been verified by numerous increasingly precise experiments, which can be expressed as: any two

point masses at the same point in a gravitational field have the same gravitational acceleration. Although this is not a surprising conclusion at all, why is this so? Two point charges in an electrostatic field are not like this. Suppose the mass of a point charge q is m , located at a place where the electric field is \vec{E} , then the electric force acting on it is $\vec{f} = q\vec{E}$, and the acceleration it acquires is

$$\vec{a} = \frac{\vec{f}}{m} = \frac{q}{m}\vec{E}. \quad (7.1.1)$$

If we place another point charge q' with a mass m' at this same point, then its acceleration will be $\vec{a}' = (q'/m')\vec{E}$. \vec{a}' and \vec{a} are not equal unless they have the same charge-to-mass ratio. When having a similar discussion about gravity, we may also distinguish the “mass” and the “charge”. The “charge” of a point mass is a measure of the amount of matter it contains, which determines the force it experiences in a gravitational field, and thus can be called the **gravitational mass**, denoted by m_G ; the “mass” of a point mass is a measure of its inertia, which determines its acceleration when a force is applied, and thus can be called the **inertial mass**, denoted by m_I [i.e., the m in (7.1.1)].² Following the discussion above it is not difficult to determine that the gravitational acceleration of a point mass in a gravitational field is $\vec{a} = (m_G/m_I)\vec{g}$, where \vec{g} is the gravitational field strength at this point. If different point masses have different mass-to-charge ratios, then they cannot have the same gravitational acceleration at the same point in a gravitational field. However, countless experiments, each one more precise than the last, have shown that the ratio m_G/m_I is the same for any point mass; by adjusting the gravitational constant G one can even set the ratio to 1 and make it as simple as $m_G = m_I$. This fact is usually called the equivalence principle (see Sect. 7.5 for details). This is an extremely unusual experimental fact which deserves serious consideration. The “charge” and “mass” for gravity are two completely different concepts, so how could they be equal? This question cannot be answered by Newton’s theory of gravity. In Newton’s theory of gravity, this is admitted as an experimental fact (it is an axiom in Newton’s formalism). Is $m_G = m_I$ just a coincidence? Could there be any deeper reason hiding underneath this fact? Could there exist a theory that is more beautiful, in which $m_G = m_I$ can be proved by reasoning? Pondering over the equivalence principle, in addition to the inspiration from Mach’s principle, led Einstein to the creation of general relativity.

The fact that $m_G = m_I$ is equivalent to the fact that all the objects in a gravitational field that experience no force other than gravity and have the same position and velocity will “march together”. This kind of characterless collective behavior strongly implies that gravity is an intrinsic property of the whole spacetime background, which

² Up to now, we have been discussing in terms of Newtonian gravity. In Newton’s theory of gravity, there are two types of gravitational mass: active and passive. The former refers to the mass of an object as a source of its gravitational field, which determines the strength of the gravitational field it produces; the latter refers to the gravitational mass of the object as a test point mass in an external gravitational field, which determines the strength of the gravitational force it experiences in a given gravitational field. The gravitational mass in the main text refers to the passive gravitational mass.

is substantially different from all the other forces. Physics is the study of the motion (evolution) of physical objects. Physical objects can be compared to actors. Just like the performance of actors cannot be done without a stage, the evolution of physical objects also always happens on some kind of stage (or background), and this stage (background) is spacetime. Before general relativity came out, people used to assume that the background spacetime of relativity is Minkowski spacetime. Minkowski spacetime is so simple that people often forgot that it exists. The “marching together” phenomenon in the gravitational field attracted Einstein’s attention to the spacetime background. Just like the actors on a lifting stage can be raised simultaneously without any effort due to the behavior of the stage itself, this “marching together” phenomenon under the gravitational force rather strongly implies that gravity is purely an effect of spacetime background. One may speculate as follows: when gravity is negligible, then the spacetime is flat; when gravity is non-negligible (e.g., when the gravitational field of the Earth or the Sun must be considered), the spacetime becomes curved, and how it is curved depends on the distribution of the matter which produces the gravitational field. According to this hypothesis, gravity is so distinct from other forces that in the 4-dimensional language it is not even a force, but the effect of curved spacetime! Therefore, a point mass that experiences no force other than gravity should be called a free point mass. Recalling that the world line of a free point mass in Minkowski spacetime is a geodesic, one can further assume naturally that the world line of a free point mass in curved spacetime is also a geodesic (of that spacetime).³ A free point mass is the simplest point mass and a geodesic is the simplest world line, and thus this assumption is also in conformity with aesthetic principles. Instead of a 4-dimensional force called “gravity” exerting on a point mass, the existence of gravity is manifested by a curved spacetime, which changes the motion of a point mass by changing its geodesic. This is the most basic postulate of general relativity. Based on this postulate, one can deduce $m_G = m_I$ as a logical consequence. (Here we come to the decisive step). Suppose two free point masses have the same initial velocity and position, i.e., their world lines intersect and their tangent vectors are equal at the intersection. Since the world line of a free point masses is a geodesic, which is uniquely determined by the initial conditions, i.e., the starting point of the geodesic and the tangent vector there (see Theorem 3.3.4), these two world lines must coincide. Translating to the language of physics, this is to say that the states of two free point masses with the same initial condition in a gravitational field must be the same at any time later, which is exactly an equivalent expression for $m_G = m_I$. Thus, once realizing that gravity in essence is the curvature of 4-dimensional spacetime, the experimental fact $m_G = m_I$, which was long mysterious in origin, is now a very natural conclusion. In its unique and elegant way, general relativity interprets gravity as a geometric effect of a 4-dimensional spacetime for the first time (which also unifies gravity and geometry for the first time), and the key to success is adding the time dimension. Solely using the 3-dimensional spacetime one cannot interpret gravity as a geometric effect.

³ The gravitational field produced by the point mass is ignored (similar to the treatment of a test charge in electromagnetism).

Remark 1 Optional Reading 8.3.2 will provide a more specific interpretation for the statement “gravity is an effect of curved spacetime” in detail.

The discussion above indicates that general relativity is a theory independent of the framework of special relativity. The framework of special relativity cannot fit general relativity or gravity.

Formulated in more modern language, the most basic postulates of general relativity can be summarized as the following three points. (The basic postulates of general relativity are summarized differently in different literature. Here is just a pedagogical way of listing them).

(a) Gravity in the 3-dimensional space in essence is the effect of the 4-dimensional spacetime curvature. That is, when gravity exists, the spacetime background is no longer Minkowski spacetime $(\mathbb{R}^4, \eta_{ab})$; instead, it is a curved spacetime (M, g_{ab}) , where M is a 4-dimensional manifold and g_{ab} is a non-flat Lorentzian metric field on M . This postulation boldly identifies gravity in physics as a pure geometric effect of the spacetime. Based on this, a point mass that experiences no force other than gravity is naturally a free point mass.

(b) The world line of a free point mass is a geodesic of the curved spacetime (M, g_{ab}) it is in. Upon postulate (a), it is pretty natural to have postulate (b). When gravity does not exist, the spacetime background is Minkowski spacetime $(\mathbb{R}^4, \eta_{ab})$. According to Sect. 6.3, the equation of motion for a point mass is

$$F^a = U^b \partial_b P^a , \quad (7.1.2)$$

where ∂_b is the derivative operator associated with the Minkowski metric η_{ab} . When gravity exists, a natural assumption is to change the ∂_b in the above equation to the derivative operator ∇_b associated with the curved metric g_{ab} , and to regard the 4-force on a free point mass as vanishing. Hence, its equation of motion is

$$0 = U^b \nabla_b (mU^a) = mU^b \nabla_b U^a , \quad (7.1.3)$$

and thus a free point mass moves along a geodesic. This proposition is very similar to the corresponding proposition without gravity, the only difference is: when gravity does not exist, the world line of a free point mass is a geodesic of Minkowski space; when gravity exists, the world line of a free point mass is a geodesic in curved spacetime. This is exactly a manifestation of the fact that general relativity is independent of special relativity. In general relativity, gravity is not represented by a 4-force on the left-hand side of the equation of motion (7.1.2), but its effect on the motion of a point mass is manifested by making the spacetime curved and requiring the point mass to move along a geodesic in the curved spacetime. In other words, the effect of gravity is substituting ∇_b for ∂_b on the right-hand side of (7.1.2).

(c) The way that the spacetime is curved is affected by the matter distribution. The specific relation is described by Einstein’s equations. [For details, see Sect. 7.7; once we have Einstein’s equations, it will be clear that (b) is not an independent postulate any more].

It can be proved that when gravity is weak enough, and the velocity of the point mass is low enough, the calculation results of general relativity agree with those of Newtonian mechanics approximately. Thus, Newtonian mechanics can be regarded as the weak-field and low-speed approximation of general relativity mechanics (see Sect. 7.8.2). Nonetheless, we should point out that, although the results are approximately the same, the viewpoints are explicitly different. Take the free fall of an apple as an example. According to Newtonian mechanics, this apple acquires an acceleration because it experiences the Earth's gravity, and thus undergoes a non-inertial motion. However, according to general relativity, the apple does not experience a 4-force, and thus is a free point mass. The effect from the Earth is that the spacetime becomes curved, and the world line of the apple is a geodesic in this curved spacetime, whose 4-acceleration (defined as $A^a \equiv U^b \nabla_b U^a$, where U^a is the 4-velocity and ∇_b is the derivative operator associated with the metric of the curved spacetime) is zero. That is, for the same motion of the apple's free fall, in Newton's theory it has a (3-dimensional) acceleration (relative to an inertial frame), while in general relativity it does not have any (4-dimensional) acceleration. Conversely, now suppose the apple is at rest on the ground. In Newton's theory, the Earth's gravity is canceled by the normal force from the ground, and thus the apple remains at rest with a zero (3-dimensional) acceleration, which undergoes an inertial motion; while in general relativity, the apple only experiences one 4-dimensional force (the normal force from the ground), and thus its world line is not a geodesic and its (4-dimensional) acceleration is nonzero. Have you realized that while you sit cosily reading this book, your 4-dimensional acceleration is not zero due to the curved spacetime caused by the Earth?

Attributing gravity to curved spacetime is a great triumph of human wisdom. Bernhard Riemann presented the concept of the intrinsic curvature as well as how to compute it when he was only 28 years old (in 1854). Before his early death (at age 40), Riemann had attempted to find a theory that unifies electromagnetism and gravity. The most important reason that it did not work out is that he focused on space and the spatial curvature rather than spacetime and the spacetime curvature. It was not until 1905 when special relativity came out that space and time were treated equally (in fact, it was not until 1908 when Hermann Minkowski brought up the absolute concept of spacetime, see Chap. 14 in Volume II). Finally, a few years after that, the groundbreaking idea that "gravity in essence is the curvature of spacetime" is gradually established along Einstein's conception of general relativity.

7.2 Physical Laws in Curved Spacetime

In the view of general relativity, every physical phenomenon is nothing but the evolution of physical objects in some curved spacetime background (M, g_{ab}). Therefore, to study physics from the viewpoint of general relativity, one first needs to find the evolution equations of those physical objects on the given curved spacetime background. Since the gravitational field in practical life or in a laboratory is too weak, the difference between general relativity and Newton's theory of gravity is normally hard to

be measured, and it is hopeless to deduce the physical laws in curved spacetime from observations or experiments. Therefore, one can only “guess” these laws by making hypotheses based on some fundamental principles, and the validity of the hypotheses can be verified by the consistency of the conclusions derived from them as well as, if possible, the results of the experiments. Of course, this “guess” is warranted, and one of the important bases is the principle of general covariance. When producing general relativity, Einstein proposed the following principle of general covariance: the mathematical expressions for all physical laws does not change under an arbitrary coordinate transformation. However, an article by E. Kretschmann in 1917 argued that this formulation for the principle of general covariance imposes no restriction on the laws of physics. Even Newton’s equation of motion can be made generally covariant by a non-substantive reformulation [see Ohanian and Ruffini (1994)]. This criticism triggered a heated discussion among physicists (including Einstein himself), and thus many different formulations for the principle of general covariance were raised. Here we introduce a formulation as follows that not only grasps the essence but is also convenient to apply [see Wald (1984) pp. 57, 68]:

Principle of General Covariance. The spacetime metric and quantities derivable from it are the only background geometric quantities that are allowed to appear in the expressions of physical laws.

Remark 1 Physical objects are like actors, while the spacetime is like the stage (background). Once a spacetime (M, g_{ab}) is given, the actors have a stage. In physical laws, there will certainly be physical quantities (dynamical quantities) that represent physical objects, such as the 4-momentum P^a of a point mass and the electromagnetic field tensor F_{ab} , etc. However, physical laws can also have spacetime geometric quantities that reflect the stage (background), which are the spacetime metric g_{ab} and quantities derivable from it (such as the derivative operator ∇_a associated with the metric g_{ab} and its $R_{abc}{}^d$, R_{ab} , R , etc.). The essence of the principle of general covariance is to eliminate all the human factors (independent of the spacetime intrinsic geometry) in the expressions of the physical law. For instance, the ordinary derivative operator ∂_a of a coordinate system or a vector field v^a assigned artificially cannot appear in a physical law, since they are neither the physical objects we study nor the intrinsic factors of the spacetime background (M, g_{ab}) . Allowing ∂_a to appear in a physical law means that the coordinate system corresponding to ∂_a is in a special position among all the coordinate systems, which is not allowed by the principle of general covariance.

Remark 2 The above formulation of the principle of general covariance is particularly suitable for textbooks adopting the abstract index notation. Many textbooks that do not use abstract indices have the following conclusion: Any physical law that can be expressed as an equality of tensors must be generally covariant. For example, suppose T and S are both tensors of type $(1, 1)$, then the equation $T = S$ is generally covariant since its component expressions in any two coordinate systems $\{x^\mu\}$ and $\{x'^\mu\}$ are obviously $T^\mu{}_\nu = S^\mu{}_\nu$ and $T'^\mu{}_\nu = S'^\mu{}_\nu$, i.e., the component expressions have the same form under any coordinate transformation (which agrees with

the formulation for the principle of general covariance by Einstein). In contrast, the Christoffel symbols $\Gamma^\sigma_{\mu\nu}$ do not obey the tensor transformation law, which means an equation that contains Christoffel symbols is not an equality of tensors, and thus is not generally covariant. However, in textbooks that use abstract indices, even the Christoffel symbol Γ^c_{ab} is regarded as a tensor (associated with a coordinate system); the same holds for $\partial_a v^b$, the result of ∂_a of a coordinate system acting on a vector field v^a . The equations that contain Γ^c_{ab} and $\partial_a v^b$ are still to be viewed as equalities of tensors. The reason why they are not generally covariant is because they do not satisfy the formulation for the principle of general covariance we introduced above, since they contain quantities not derivable from g_{ab} , i.e., Γ^c_{ab} and $\partial_a v^b$, which puts the coordinate system corresponding to Γ^c_{ab} and $\partial_a v^b$ in a special position. In a word, both kinds of textbooks say that an equation containing Christoffel symbols is not generally covariant, but their reasons are different (due to different formulations of the principle of general covariance).

Based on the discussion above, we can put forward two principles that the physical laws in curved spacetime must obey: (a) the principle of general covariance; (b) when g_{ab} equals the Minkowski metric η_{ab} , they should go back to the physical laws in special relativity.⁴ Although these two necessary criteria cannot uniquely determine the physical laws in curved spacetime, one can use them as guidance, together with physical and aesthetic considerations, to acquire the physical laws naturally in many cases. Since the difference between general relativity and special relativity is nothing but the difference between the spacetime background [i.e., between (M, g_{ab}) and $(\mathbb{R}^4, \eta_{ab})$], the 4-dimensional description of physical objects in special relativity can be naturally generalized to general relativity. For instance, the world lines of point masses and photons are still timelike and null curves, respectively (of course, this actually already generalizes the connotation of “the principle of invariant light speed” and “point masses must move slower than light” to general relativity); the proper time of a point mass is still the length of its world line, the 4-velocity U^a of a point mass is still defined as the unit tangent vector of its world line, and the 4 momentum is still defined as $P^a := mU^a$ (m is the rest mass); the energy of a point mass relative to an instantaneous observer (p, Z^a) is still defined as $E := -P^a Z_a$, and an electromagnetic field is still described by a 2-form field F_{ab} , etc. In order to find the physical laws obeyed by these physical quantities, in most of the cases one only needs to substitute all the η_{ab} and ∂_a in the expressions for the corresponding laws in special relativity with g_{ab} and ∇_a . This method may be dubbed the “**minimal substitution rule**”. It is easy to see that a formula obtained in this manner obeys the two principles we stated above. Here are some examples of applying this rule: the 4-acceleration of a point mass in curved spacetime is defined as

$$A^a := U^b \nabla_b U^a, \quad (7.2.1)$$

⁴ Principle (a) is put in the same way in all textbooks (although the formulation for the principle of general covariance may be different); however, there are at least two ways of stating the principle (b) in different books. The other one is: (b) the equivalence principle. With regard to the effects of the physical laws being derived, these two ways are equivalent. For details, see Sect. 7.5.

and the 4-force exerting on the point mass is defined as

$$F^a := U^b \nabla_b P^a . \quad (7.2.2)$$

For a free point mass, $F^a = 0$ (gravity is not a 4-force!), and the equation above becomes $U^b \nabla_b U^a = 0$, i.e., the geodesic equation, which agrees with the basic postulate (b) of general relativity (see Sect. 7.1). For a point mass in an electromagnetic field, its equation of motion is then

$$q F^a_b U^b = U^b \nabla_b P^a . \quad (7.2.3)$$

Note that the effect from the electromagnetic field F_{ab} on the point mass is manifested on the left-hand side of the equation (as a 4-force $q F^a_b U^b$), while the effect from gravity on the point mass is manifested on the right-hand side of the equation (by the derivative ∇_a not being ∂_a). The equations of motion of the electromagnetic field F_{ab} (Maxwell's equations in curved spacetime) should be

$$\nabla^a F_{ab} = -4\pi J_b , \quad (7.2.4)$$

$$\nabla_{[a} F_{bc]} = 0 . \quad (7.2.5)$$

The energy-momentum tensor of the electromagnetic field should be expressed as

$$T_{ab} = \frac{1}{4\pi} (F_{ac} F_b{}^c - \frac{1}{4} g_{ab} F_{cd} F^{cd}) . \quad (7.2.6)$$

Another important basis for this equation holding in curved spacetime is that it satisfies $\nabla^a T_{ab} = -F_{bc} J^c$ [see Exercise 6.18 (a)], which indicates that the total energy, momentum and angular momentum of the electromagnetic field and charged particle field are all conserved (see the end of Sect. 6.6.4). The reader should verify this equation.

Since (7.2.5) can be expressed as $d\mathbf{F} = 0$, we can at least locally introduce an electromagnetic 4-potential \mathbf{A} such that $\mathbf{F} = d\mathbf{A}$, and hence (7.2.4) can be expressed in terms of \mathbf{A} as

$$-4\pi J_b = \nabla^a (\nabla_a A_b - \nabla_b A_a) = \nabla^a \nabla_a A_b - \nabla^a \nabla_b A_a . \quad (7.2.7)$$

In special relativity, the second term on the right-hand side of the equation above is $-\partial^a \partial_b A_a$, which can be easily rewritten as $-\partial_b \partial^a A_a$, and then using the Lorenz gauge condition we can express (7.2.7) in special relativity as

$$\partial^a \partial_a A_b = -4\pi J_b \quad [\text{cf. (6.6.31)}] .$$

However, now ∇_a and ∇_b do not commute, if we want to use the Lorenz condition $\nabla^a A_a = 0$ we need to rewrite the second term on the right-hand side of (7.2.7) using (3.4.4) as $-\nabla^a \nabla_b A_a = -\nabla_b \nabla^a A_a - R_b{}^d A_d = -R_b{}^d A_d$, which turns (7.2.7) into

$$\nabla^a \nabla_a A_b - R_b{}^d A_d = -4\pi J_b. \quad (7.2.8)$$

Interestingly, if we use the minimal substitution rule directly to the equation (6.6.31) in special relativity, we have

$$\nabla^a \nabla_a A_b = -4\pi J_b, \quad (7.2.9)$$

which is obviously different from (7.2.8). This example indicates that the minimal substitution rule does not uniquely determine the physical laws in some circumstances. More consideration needs to be taken when cases like this are encountered. For this example, it can be shown that (7.2.8) leads to the law of charge conservation $\nabla_a J^a = 0$ (Exercise 7.1) while (7.2.9) does not. From this physical consideration, we choose (7.2.8) as the equation of motion of the 4-potential A . The ambiguity of this example comes from the non-commutativity of the derivative operators, which is a problem that all the equations containing second or higher derivatives (with two or more ∇_a acting successively) will encounter when transferred from special relativity to general relativity. The reader may compare this with the following fact: When transferred from classical mechanics to quantum mechanics, the non-commutativity of the operators is also the source of ambiguity.

[Optional Reading 7.2.1]

For a source-free electromagnetic field, (7.2.8) becomes

$$\nabla^a \nabla_a A_b - R_b{}^d A_d = 0. \quad (7.2.8')$$

Inspired by the discussion at the end of Sect. 6.6.5 (before Optional Reading 6.6.5), we want to consider a wave solution $A_b = C_b \cos \theta$ of the equation above, which is a product of the “slowly changing” amplitude C_b and the “rapidly changing” phase factor, and look for the possibility of applying the geometric optics approximation. The difference between (7.2.8') and the corresponding equation $\partial^a \partial_a A_b = 0$ in Minkowski spacetime is that the former contains the curvature term $R_b{}^d A_d$, which needs to be negligible if we want to apply the geometric optics approximation. Consider three length scales as follows:

- (1) The characteristic length \tilde{L} above which the change of C_b or $K^a \equiv \nabla^a \theta$ is notable;
- (2) The length that describes the “magnitude” of the spacetime curvature

$$\tilde{R} \equiv |R_{\mu\nu\rho\sigma}|^{-1/2},$$

where $R_{\mu\nu\rho\sigma}$ is a typical component of R_{abcd} in a typical local inertial frame (see Sect. 7.5 for details);

- (3) The wavelength λ ($\lambda \equiv 2\pi/\omega$, $\omega \equiv -Z^a K_a$) of A_b relative to the local inertial frame we mentioned above.

If these three satisfy $\lambda \ll \tilde{L}$ and $\lambda \ll \tilde{R}$, then both the derivative term $\nabla^a \nabla_a C_b$ and the curvature term $R_b{}^d A_d$ can be neglected, and thus we have approximately

$$(\nabla^a \theta) \nabla_a \theta = 0. \quad (7.2.10)$$

Hence, $K^a \equiv \nabla^a \theta$ is still the null normal vector of the null hypersurface $\mathcal{S} = \{p \in \mathbb{R}^4 \mid \theta_p = C\}$ ($C = \text{constant}$), the integral curves of K^a are still null geodesics (the proof is similar to Sect. 6.6.5, note that ∇_a being torsion free assures that $\nabla_a \nabla_b \theta = \nabla_b \nabla_a \theta$), a light signal still

propagates along a null geodesic, and the angular frequency of the electromagnetic wave (photon) relative to an observer with a 4-velocity Z^a is still

$$\omega = -K_a Z^a, \quad (7.2.11)$$

and so on. Thus, the geometric optics approximately holds when $\lambda \ll \tilde{L}$ and $\lambda \ll \tilde{R}$. This approximation is used in many places in this text (such as Sect. 9.2.1 and Sect. 10.2.2).

References for the geometric optics in curved spacetime are: Wald (1984) p. 71; Misner et al. (1973) Sect. 22.5; Straumann (1984) pp. 100–103.

[The End of Optional Reading 7.2.1]

[Optional Reading 7.2.2]

Maxwell's equations (7.2.4) and (7.2.5) in curved spacetime also have the following formulation in terms of the exterior differentiation operator:

$$d^* F = 4\pi^* J, \quad (7.2.4')$$

$$dF = 0, \quad (7.2.5')$$

where ${}^* F$ is the dual form of $F \equiv F_{ab}$ (see Sect. 5.6), which is still a 2-form, and ${}^* J$ is the dual 3-form of the 1-form J_a . The equivalence of (7.2.5') and (7.2.5) can be seen directly from the definition of exterior differentiation, while the equivalence of (7.2.4') and (7.2.4) is a bit more tricky to show. By definition, $(d^* F)_{fab} = d_f(\varepsilon_{abcd} F^{cd}/2) = 3\nabla_{[f}(\varepsilon_{ab]cd} F^{cd})/2$. Contracting the right-hand side with ε^{efab} yields $3\varepsilon^{efab}\varepsilon_{cdab}(\nabla_f F^{cd})/2 = -3 \times 4\varepsilon_c{}^e \delta_d{}^f (\nabla_f F^{cd})/2 = -6\nabla_f F^{ef}$, and so $\varepsilon^{efab}(d^* F)_{fab} = 6\nabla_f F^{fe}$. Contracting this equation again with ε_{egcd} yields $-(d^* F)_{gcd} = \varepsilon_{egcd}\nabla_f F^{fe}$. It is not difficult to see from the definition ${}^* J_{gcd} \equiv J^e \varepsilon_{egcd}$ that the above equation can be expressed as (7.2.4') if and only if (7.2.4) holds. Thus, (7.2.4') and (7.2.4) are equivalent.

[The End of Optional Reading 7.2.2]

7.3 Fermi-Walker Transport and Non-Rotating Observers

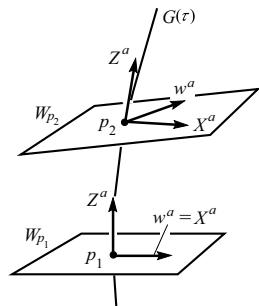
After reading Sect. 7.1, many readers may want to learn more about topics like the equivalence principles, Einstein's elevator, local inertial frames, and the relationship between gravity and an inertial force. To have a precise understanding of these topics some basic concepts will be necessary. This section will introduce an important one, namely the concept of a non-rotating observer. (The observer in an Einstein elevator is not only free-falling, but also non-rotating).

Imagine you are traveling around the world on an airplane. A small arrow is fixed in front of you, perpendicular to your chest and pointing away from you. At a proper time τ_1 you take a nap, and until you wake up at τ_2 , the arrow will of course still be perpendicular to your chest, but the spatially-pointed direction can be different from that at τ_1 since the motion of the airplane is arbitrary. If the pointed direction has changed, it is natural to say that the arrow “changed its direction” in $\Delta\tau \equiv \tau_2 - \tau_1$, or it rotated in $\Delta\tau$. However, what does it mean by “changing its direction”? How do we judge if the direction is changed or not? This is actually to ask: what is a rotation? How to determine if a rotation occurs? The answer is clear in Newtonian mechanics:

the axis of a gyroscope flywheel (or “a gyroscope axis” for short) represents a fixed direction [see Sachs and Wu (1977) pp. 50, 52]. If you have a gyroscope in your hand, the arrow and the gyroscope axis are parallel at τ_1 but not parallel at τ_2 , then we can conclude that the arrow has rotated within $\Delta\tau$. This criterion can be generalized to general relativity.

Now we translate this criterion into the 4-dimensional language. Let $G(\tau)$ represent your world line, then at the time τ_1 the arrow is represented by a spatial vector w^a at a point $p_1 \equiv G(\tau_1)$ (“spatial” means it is perpendicular to your 4-velocity $Z^a|_{p_1}$ at τ_1). For convenience’s sake, we set the magnitude of w^a to 1. As your proper time flows, the arrow corresponds to a spatial vector field with unit length on the curve $G(\tau)$. Similarly, if we also represent the direction of the gyroscope axis at each time using a unit vector, then the gyroscope axis corresponds to another spatial vector field X^a with unit length on $G(\tau)$. The 3-dimensional description we mentioned before indicates that w^a and X^a coincide at $p_1 \equiv G(\tau_1)$ but do not coincide at $p_2 \equiv G(\tau_2)$ (see Fig. 7.1). Since we stipulate X^a to represent the non-rotating direction, we say that w^a rotated in $\Delta\tau \equiv \tau_2 - \tau_1$. To describe the rotating vector field w^a on the world line $G(\tau)$, we should first describe the non-rotating vector field X^a , since it is the criterion for measuring the rotation of w^a . As a non-rotating spatial vector field on $G(\tau)$, what mathematical property does X^a have? A natural guess is: X^a is a vector field parallelly transported along $G(\tau)$. However, except for special cases, this is not a correct guess. The key point is that the vector field parallelly transported along $G(\tau)$ determined by a spatial vector $X^a|_{p_1}$ at $p_1 \equiv G(\tau_1)$ is not a spatial vector field in general. [Proof: suppose X^a is parallelly transported along $G(\tau)$, then $Z^b \nabla_b (X^a Z_a) = X^a Z^b \nabla_b Z_a = X^a A_a$, where ∇_a is the derivative operator associated with the spacetime metric g_{ab} , and A^a is the 4-acceleration of $G(\tau)$. As long as $G(\tau)$ is not a geodesic, and X^a is not orthogonal to A^a , then the right-hand side of the above equation is nonzero. Hence, $X^a Z_a$ is not a constant along $G(\tau)$, and cannot be everywhere vanishing on $G(\tau)$]. To describe the motion of a non-rotating spatial vector field X^a along $G(\tau)$, E. Fermi (in 1922) and A. G. Walker (in 1923) introduced a derivative notion along a curve, which is of physical importance and closely related to, but different from, a covariant derivative. This derivative, dubbed the Fermi-Walker derivative, is defined as follows:

Fig. 7.1 X^a and w^a are both spatial vector fields on $G(\tau)$. X^a represents the gyroscope axis. The spatial rotation of w^a can be seen by comparing with X^a



Definition 1 Suppose $G(\tau)$ is a timelike curve⁵ (where τ is the proper time) in the spacetime (M, g_{ab}) , and $\mathcal{F}_G(k, l)$ ⁶ represents the collection of all smooth tensor fields of type (k, l) along $G(\tau)$. A map $D_F/d\tau : \mathcal{F}_G(k, l) \rightarrow \mathcal{F}_G(k, l)$ is called a **Fermi-Walker derivative operator** (or Fermi derivative for short) if it satisfies the following conditions:

- (a) Linearity ;
- (b) Leibniz rule ;
- (c) Commutativity with contraction ;

$$(d) \frac{D_F f}{d\tau} = \frac{df}{d\tau} \quad \forall f \in \mathcal{F}_G(0, 0); \quad (7.3.1)$$

$$(e) \frac{D_F v^a}{d\tau} = \frac{Dv^a}{d\tau} + (A^a Z^b - Z^a A^b)v_b \quad \forall v^a \in \mathcal{F}_G(1, 0), \quad (7.3.2)$$

where $Z^a \equiv (\partial/\partial\tau)^a$ represents the 4-velocity of $G(\tau)$, $A^a \equiv Z^b \nabla_b Z^a$ represents the 4-acceleration of $G(\tau)$, and $Dv^a/d\tau$ is another notation for $Z^b \nabla_b v^a$, the covariant derivative along $G(\tau)$ (where ∇_b satisfies $\nabla_b g_{ac} = 0$).

Remark 1 Condition (e) stipulates the expression for the Fermi derivative of a vector field, and combining it with the other conditions yields the results of $D_F/d\tau$ acting on an arbitrary tensor field.

Proposition 7.3.1 *The Fermi derivative has the following properties:*

- (1) *If $G(\tau)$ is a geodesic, then $D_F v^a/d\tau = Dv^a/d\tau$;*
- (2) $D_F Z^a/d\tau = 0$;
- (3) *If w^a is a spatial vector field on $G(\tau)$ ($w^a Z_a = 0$ for each point on the world line), then*

$$D_F w^a/d\tau = h^a{}_b (Dw^b/d\tau), \quad (7.3.3)$$

where $h_{ab} = g_{ab} + Z_a Z_b$, and $h^a{}_b = g^{ac} h_{cb}$ is the projection map at each point of $G(\tau)$. This property guarantees that the Fermi derivative of a spatial vector field is still a spatial vector field.

- (4) $D_F g_{ab}/d\tau = 0$, and equivalently

$$D_F(g_{ab} v^a u^b)/d\tau = g_{ab} v^a D_F u^b/d\tau + g_{ab} u^b D_F v^a/d\tau \quad \forall v^a, u^a \in \mathcal{F}_G(1, 0). \quad (7.3.4)$$

⁵ We only discuss the case where $G(\tau)$ is a non-self-intersecting curve, otherwise one will encounter causal difficulties (see Chap. 11 in Volume II). In fact, the timelike curves representing observers in this text are all assumed to be non-self-intersecting curves.

⁶ Note that we are abusing the notation here, since $\mathcal{F}_M(k, l)$ technically denotes the collection of all the tensor fields of type (k, l) on the manifold M but some fields in $\mathcal{F}_G(k, l)$ here do not lie on the curve G .

Proof Property (1) can be easily seen from (7.3.2). Property (2) can be easily proved from (7.3.2) and the definition of A^a (using $A^a Z_a = 0$). The proof for property (3) is left as Exercise 7.3. The proof for property (4) is as follows:

$$\begin{aligned} g_{ab} v^a D_F u^b / d\tau + g_{ab} u^b D_F v^a / d\tau &= v_a D_F u^a / d\tau + u_a D_F v^a / d\tau \\ &= v_a (Du^a / d\tau + 2A^{[a} Z^{b]} u_b) + u_a (Dv^a / d\tau + 2A^{[a} Z^{b]} v_b) \\ &= v_a Du^a / d\tau + u_a Dv^a / d\tau + 4A^{[a} Z^{b]} v_{(a} u_{b)} = D(v_a u^a) / d\tau = D_F(g_{ab} v^a u^b) / d\tau, \end{aligned}$$

where in the last step we used (7.3.1). \square

Definition 2 A vector field v^a is said to be **Fermi-Walker transported** along $G(\tau)$ if

$$\frac{D_F v^a}{d\tau} = 0.$$

Fermi-Walker transport is also called **Fermi transport** for short.

Remark 2 Property (1) of the Fermi derivative indicates that Fermi transport along a geodesic is parallel transport; property (2) indicates that the 4-velocity of $G(\tau)$ is always Fermi transported along $G(\tau)$; from property (4) we can see that $D_F v^a / d\tau = 0 = D_F u^a / d\tau \Rightarrow d(g_{ab} v^a u^b) / d\tau = 0$, which can be abbreviated as “Fermi transport preserves the inner product”, similar to “parallel transport preserves the inner product”.

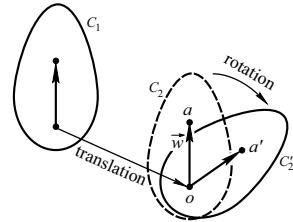
Proposition 7.3.2 $p \in G$ and $v^a \in V_p$ uniquely determine a vector field Fermi transported along $G(\tau)$.

Proof Omitted. [The reader may refer to Sachs and Wu (1977) p. 51 and the reference therein]. \square

Remark 3 ① From the fact that Z^a is Fermi transported along $G(\tau)$ and that Fermi transport preserves the inner product, we can see that the vector field v^a Fermi transported along $G(\tau)$ determined by a spatial vector $v^a|_p \in V_p$ is everywhere perpendicular to Z^a , and hence is a spatial vector field. ② Each basis vector of an orthonormal tetrad (whose zeroth basis vector equals $Z^a|_p$) at $p \in G$ determines a vector field Fermi transported along $G(\tau)$ based on Proposition 7.3.2, and from the fact that Fermi transport preserves the inner product we can see that these four vector fields are orthonormal at each point of the curve. Thus, an orthonormal tetrad at p uniquely determines an orthonormal tetrad field Fermi transported along $G(\tau)$, in which the zeroth basis vector field is the tangent vector field Z^a along $G(\tau)$.

Fermi transport has an important physical meaning: the necessary and sufficient condition for a spatial vector field w^a with a constant magnitude on a world line $G(\tau)$ to have no spatial rotation is that w^a is Fermi transported along $G(\tau)$, i.e., $D_F w^a / d\tau = 0$ (for the reason see Proposition 7.3.6). Therefore, a gyroscope axis (which can be viewed as a unit vector) is a spatial vector field Fermi transported along

Fig. 7.2 A rigid motion can be decomposed into a translation and a rotation



the world line of the gyroscope. For instance, suppose $\{t, x, y, z\}$ is a Lorentzian system of Minkowski spacetime, and $G(\tau)$ is a t -coordinate line of this system, then the coordinate basis vectors $(\partial/\partial t)^a, (\partial/\partial x)^a, (\partial/\partial y)^a, (\partial/\partial z)^a$ are all Fermi transported along $G(\tau)$, and thus the latter three are non-rotating spatial vector fields on $G(\tau)$, which physically represent the three axes of the gyroscope (orthogonal to each other). Conversely, if a spatial vector w^a (with a constant magnitude) is not Fermi transported along $G(\tau)$, then it has a spatial rotation.

In order to introduce Proposition 7.3.6, we first talk about the definition of a spatial rotation. In Newtonian mechanics, any motion of a rigid body can be decomposed into a translation and a rotation. Figure 7.2 represents the motion of a rigid body from a configuration C_1 to another configuration C'_2 . This can be done in two steps: first move to a configuration C_2 by a translation, and then arrive at C'_2 by a rotation with respect to a fixed point o (called the “base point”). To describe this rotation, one can choose another point of the rigid body, whose position turns from a to a' during the rotation. Just as the motion of the base point represents the translation of the body, the motion of the point a (from a to a') represents the rotation of the body. Let $\vec{w}(t)$ be the position vector of a relative to o , then the rotation of the rigid body is manifested by $d\vec{w}(t)/dt \neq 0$, and thus can be described by the rotation of the vector $\vec{w}(t)$. More precisely, the vector $\vec{w}(t)$ with one end fixed at o is said to be **rotating** if there exists a vector $\vec{\omega}(t)$ such that

$$\frac{d\vec{w}(t)}{dt} = \vec{\omega}(t) \times \vec{w}(t), \quad (7.3.5)$$

where $\vec{\omega}(t)$ is called the **(instantaneous) angular velocity** of the rotation. Noticing that $d(\vec{w} \cdot \vec{w})/dt = 2\vec{w} \cdot d\vec{w}/dt = 2\vec{w} \cdot (\vec{\omega} \times \vec{w}) = 0$, we can see that a rotation preserves the magnitude of a vector. From the above definition of a vector’s rotation, one can prove using Newtonian mechanics that a gyroscope axis (as a unit vector) is non-rotating, i.e., its $\vec{\omega} = 0$. Hence, a gyroscope axis represents a non-rotating direction.

To generalize the Newtonian definition above for a rotation of a vector to special relativity (and then to general relativity), we first rewrite (7.3.5) in terms of the components in a Cartesian system (or physically called a Galilean system) as

$$\frac{dw^i(t)}{dt} = \varepsilon^i_{jk} \omega^j w^k, \quad (7.3.5')$$

and imagine that there is an observer G at the base point o (the end of $\vec{\omega}$). Since o is at rest relative to an inertial frame, the world line $G(\tau)$ of G should be a geodesic when carried over to special relativity, and \vec{w} is a spatial vector field w^a on the curve. Let $\{t, x^i\}$ represent the coordinates of the observer G 's inertial frame, then on $G(\tau)$ we have $t = \tau$. Hence, we have the following generalization for the definition of a rotation in special relativity: a spatial vector field $w^a(\tau)$ on a timelike geodesic $G(\tau)$ in Minkowski spacetime is said to be **rotating** if there exists a spatial vector field $\omega^a(\tau)$ on $G(\tau)$ such that

$$\frac{dw^i(\tau)}{d\tau} = \varepsilon^i_{jk}\omega^j w^k, \quad (7.3.6)$$

where w^i and ω^j are the i th and j th components of w^a and ω^a , respectively, in the system $\{t, x^i\}$. For any point p on $G(\tau)$, if we lower the index of the angular velocity vector ω^a and make it an angular velocity 1-form ω_a using the induced metric h_{ab} of W_p , and use Ω_{ab} to represent the dual differential form of ω_a in W_p , i.e., $\Omega_{ab} \equiv (*\omega)_{ab} = \omega^c \varepsilon_{cab}$ (where ε_{cab} is the volume element associated with h_{ab}), then Ω_{ab} is called the **angular velocity 2-form**, using which one can rewrite (7.3.6) as

$$\frac{dw^i}{d\tau} = -\Omega^{ij} w_j. \quad (7.3.7)$$

Take an orthonormal spatial triad field $\{(e_i)^a\}$ on the world line such that $(e_3)^a$ is parallel to ω^a , then $\omega^1 = \omega^2 = 0$, $\omega^3 \neq 0$, and so we can say that w^a is rotating with respect to the axis $(e_3)^a$. On the other hand, from $\Omega_{ab} = \omega^c \varepsilon_{cab}$ we know that $\{\omega^1 = \omega^2 = 0, \omega^3 \neq 0\}$ corresponds to $\{\Omega_{23} = \Omega_{31} = 0, \Omega_{12} \neq 0\}$, and hence one can also say that ω^a is rotating in the $(1, 2)$ -plane (generally, a rotation in the (i, j) -plane means that the nonzero components of Ω_{ab} are Ω_{ij} and Ω_{ji}). These two statements are equivalent for a 3-dimensional vector space W_p , but the latter one is more convenient to be carried over to 4 dimensions. Now, it is not necessary to restrict the spatial rotation of a spatial vector field on a geodesic in Minkowski spacetime. Here we will generalize the definition for the “spacetime rotation” of an arbitrary vector field on an arbitrary timelike curve in any spacetime.

Definition 3 Suppose $G(\tau)$ is the world line (not necessarily a geodesic) of an arbitrary observer in the spacetime (M, g_{ab}) , and v^a is a vector field on $G(\tau)$ (not necessarily a spatial vector field). If there exists a 2-form field Ω_{ab} on $G(\tau)$ such that

$$\frac{Dv^a}{d\tau} = -\Omega^{ab} v_b, \quad (7.3.8)$$

then we say that v^a undergoes a **spacetime rotation** with an angular velocity Ω_{ab} . In other words, the angular velocity 2-form for the spacetime rotation of v^a is Ω_{ab} . If $Dv^a/d\tau = 0$, then we say v^a has no spacetime rotation.

Proposition 7.3.3 Suppose two vector fields v^a and u^a on $G(\tau)$ undergo the same spacetime rotation Ω_{ab} , then $v^a u_a$ is a constant on $G(\tau)$.

Proof

$$\frac{D}{d\tau}(v^a u_a) = u_a \frac{Dv^a}{d\tau} + v_a \frac{Du^a}{d\tau} = u_a(-\Omega^{ab} v_b) + v_a(-\Omega^{ab} u_b) = -2\Omega^{ab} v_{(a} u_{b)} = 0,$$

where the antisymmetry of Ω_{ab} is used in the last step. \square

Proposition 7.3.3 indicates that a spacetime rotation preserves the magnitude of a vector (which can be easily seen by taking $v^a = u^a$), and thus only a vector field v^a with a constant magnitude along $G(\tau)$ can be a vector field undergoing spacetime rotation. Conversely, one can show that (Exercise 7.4) a vector field v^a with a (nonvanishing) constant magnitude on $G(\tau)$ must undergo a spacetime rotation.

Remark 4 Suppose $\vec{\omega}$ satisfies (7.3.5), and a spatial vector $\vec{\lambda}$ satisfies $\vec{\lambda} \times \vec{w} = 0$ (i.e., there exists a coefficient β such that $\vec{\lambda} = \beta \vec{w}$), then $\vec{\omega}' \equiv \vec{\omega} + \vec{\lambda}$ also satisfies (7.3.5). This reflects nothing but the following fact: no matter how \vec{w} rotates, one can always add to this rotation an additional arbitrary rotation $\vec{\lambda} = \beta \vec{w}$ with respect to \vec{w} itself, since a vector “rotating with respect to itself” is the same as non-rotating. Similarly, suppose Ω_{ab} satisfies (7.3.8), and a 2-form Λ_{ab} satisfies $\Lambda^{ab} v_b = 0$, then $\Omega'_{ab} = \Omega_{ab} + \Lambda_{ab}$ also satisfies (7.3.8). This Λ_{ab} reflects the “gauge freedom” of Ω_{ab} , i.e., to v^a there is essentially no difference if two Ω_{ab} only differ by a Λ_{ab} satisfying $\Lambda^{ab} v_b = 0$. One can choose the most convenient one among these Ω_{ab} in a discussion (see Optional Reading 7.3.1). For instance, according to Definition 3, one can say that a necessary and sufficient condition for v^a to have no spacetime rotation is that its $\Omega_{ab} = 0$, although there also exist many choices of nonvanishing Ω_{ab} that satisfy $Dv^a/d\tau = 0$. (Thus, this “necessary and sufficient condition” can differ by a gauge transformation. The same for some other “necessary and sufficient conditions” in this section).

A non-geodesic world line has $DZ^a/d\tau \neq 0$, and hence its Z^a undergoes a spacetime rotation. We now would like to find the angular velocity for this spacetime rotation.

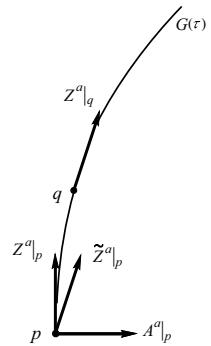
Proposition 7.3.4 *The angular velocity 2-form for the spacetime rotation of the 4-velocity Z^a of $G(\tau)$ is $\tilde{\Omega}_{ab} = A_a \wedge Z_b$, where A^a is the 4-acceleration of $G(\tau)$.*

Proof $-\tilde{\Omega}^{ab} Z_b = -(A^a Z^b - Z^a A^b) Z_b = A^a = DZ^a/d\tau$, comparing with (7.3.8) we can see that the angular velocity 2-form for the spacetime rotation of Z^a is $\tilde{\Omega}_{ab}$. \square

From $\tilde{\Omega}_{ab} = A_a \wedge Z_b$ we can see that the spacetime rotation represented by $\tilde{\Omega}_{ab}$ takes place in the (Z^a, A^a) -plane, whose spatial components in the orthonormal tetrad with Z^a as $(e_0)^a$ are $\tilde{\Omega}_{ij} = 0$. Such a spacetime rotation is called a **pseudo-rotation** [see Misner et al. (1973) p. 170]. Conversely, a spacetime rotation Ω_{ab} that only has spatial components (i.e., $\Omega_{0i} = 0$, $i = 1, 2, 3$) is called a (pure) **spatial rotation**, or simply a **rotation** when there is no confusion.

It follows from $\tilde{\Omega}_{ab} = A_a \wedge Z_b$ that the 4-acceleration $A^a \neq 0$ [i.e., $G(\tau)$ deviates from a geodesic] is the basic reason (necessary and sufficient condition) for Z^a to

Fig. 7.3 The cause of a pseudo-rotation: $G(\tau)$ being non-geodesic forces the 4-velocity Z^a to “rotate” in the (Z^a, A^a) -plane during the transition from p to q



undergo a pseudo-rotation. This can be explained intuitively by means of Fig. 7.3. According to Theorem 3.2.4, we can see from $A^a = Z^b \nabla_b Z^a$ that

$$A^a|_p = \lim_{\Delta\tau \rightarrow 0} \frac{1}{\Delta\tau} (\tilde{Z}^a|_p - Z^a|_p),$$

where $\tilde{Z}^a|_p$ is the result of $Z^a|_p$ parallel transported along $G(\tau)$ to p , and $\Delta\tau \equiv \tau(q) - \tau(p)$. It can be intuitively seen from Fig. 7.3 that: ① The deviation of $G(\tau)$ from a geodesic forces its tangent Z^a to “rotate” (pseudo-rotate) during the transition from p to q ; ② This pseudo-rotation is indeed in the (Z^a, A^a) -plane.

Since any spatial vector field w^a on $G(\tau)$ is orthogonal to Z^a , the fact that Z^a undergoes a pseudo-rotation $\hat{\Omega}_{ab}$ also forces w^a to undergo such a pseudo-rotation. Now we will prove that subtracting this inevitable pseudo-rotation from the spacetime rotation of w^a must yield a pure spatial rotation.

Proposition 7.3.5 Suppose $\tilde{\Omega}_{ab}$ is the pseudo-rotation experienced by the 4-velocity Z^a of $G(\tau)$, and Ω_{ab} is the spacetime rotation experienced by a spatial vector field $w^a (\neq 0)$ on $G(\tau)$, then $\hat{\Omega}_{ab} \equiv \Omega_{ab} - \tilde{\Omega}_{ab}$ is a pure spatial rotation (which may differ by a gauge transformation).

Proof See Optional Reading 7.3.1. □

Proposition 7.3.6 The necessary and sufficient condition for a spatial vector field w^a with a constant magnitude on the world line $G(\tau)$ of an observer to have no spatial rotation is that w^a is Fermi transported along $G(\tau)$, i.e., $D_F w^a / d\tau = 0$.

Proof Since w^a has a constant magnitude, from the paragraph above Remark 4 we know that w^a undergoes a spacetime rotation, i.e., there exists an Ω_{ab} such that $Dw^a / d\tau = -\Omega^{ab} w_b$. Combining this with $\hat{\Omega}_{ab} \equiv \Omega_{ab} - \tilde{\Omega}_{ab}$ yields

$$-\hat{\Omega}^{ab} w_b = \frac{Dw^a}{d\tau} + \tilde{\Omega}^{ab} w_b. \quad (7.3.9)$$

Noticing that $\tilde{\Omega}^{ab} = A^a Z^b - Z^a A^b$, the equation above can also be expressed as

$$\frac{D_F w^a}{d\tau} = -\hat{\Omega}^{ab} w_b . \quad (7.3.10)$$

Since $\hat{\Omega}_{ab}$ represents the spatial rotation of w^a , the necessary and sufficient condition for a spatial vector field w^a with a constant magnitude to have no spatial rotation is that $D_F w^a / d\tau = 0$. \square

Conversely, suppose w^a has a spatial rotation, let ω_a be the dual form of $\hat{\Omega}_{ab}$ (the Hodge dual in the 3-dimensional space W_p of $p \in G$), i.e.,

$$\hat{\Omega}_{ab} = \omega^c \varepsilon_{cab} . \quad (7.3.11)$$

Equation (7.3.10) can then be rewritten as

$$\frac{D_F w^a}{d\tau} = -\varepsilon^a{}_{bc} w^b \omega^c . \quad (7.3.12)$$

Or, let ε_{abcd} represent the volume element associated with g_{ab} , then (7.3.12) can also be written using $\varepsilon_{bcd} = Z^a \varepsilon_{abcd}$ as

$$g_{ab} \frac{D_F w^b}{d\tau} = \varepsilon_{abcd} Z^b w^c \omega^d . \quad (7.3.12')$$

The ω_a defined by (7.3.11) is called the **spatial angular velocity** (or **angular velocity** for short) of the spatial vector field w^a . That is, a non-Fermi transported spatial vector field w^a can be described by a nonzero spatial angular velocity ω^a .

Suppose $\{(e_i)^a\}$ is an orthonormal spatial triad field on $G(\tau)$. Since any two basis vectors are orthogonal, they have a “rigid relationship”, and one can expect that these three basis vectors have the same spacetime angular velocity Ω_{ab} , and thus have the same spatial angular velocity $\hat{\Omega}_{ab}$. See the following proposition:

Proposition 7.3.7 *The three basis vector fields in any orthonormal spatial triad field $\{(e_i)^a\}$ on $G(\tau)$ have the same spatial angular velocity $\hat{\Omega}_{ab}$ (no more gauge freedom).*

Proof See Optional Reading 7.3.1. \square

Remark 5 ① This $\hat{\Omega}_{ab}$ shared by each $(e_i)^a$ is called the angular velocity 2-form for the spatial rotation of this triad field, and the corresponding ω^a (satisfying $\hat{\Omega}_{ab} = \omega^c \varepsilon_{cab}$) is called the **spatial angular velocity vector** of this triad field. ② One may ask: suppose $(e_1)^a$ and $(e_2)^a$ rotates with respect to $(e_3)^a$ with an angular velocity ω^a [parallel to $(e_3)^a$], then $(e_3)^a$ is non-rotating, and hence has zero angular velocity. How can one say that these three vectors have the same angular velocity? The answer is: using the “gauge freedom” (see Remark 4), one can say that the angular velocity of $(e_3)^a$ is also ω^a (since a rotation with respect to itself is

equivalent to no rotation), and so there is no contradiction. Thus, we can also see that the proof of Proposition 7.3.7 requires the use of the gauge freedom. It should be emphasized that: when one finds that a basis vector in a spatial triad field [e.g., $(e_3)^a$] is non-rotating along a curve, one cannot assert based on Proposition 7.3.7 that the other two basis vector are also non-rotating, since they can rotate with respect to $(e_3)^a$.

Remark 6 The discussion above indicates that an observer is determined by two factors: ① a world line $G(\tau)$, and ② an orthonormal tetrad field on $G(\tau)$ [which satisfies $(e_0)^a = Z^a$]. In some cases factor ② is not critical, so one only needs to specify the world line $G(\tau)$ when talking about an observer. Therefore, some authors treat an observer as a world line [e.g., Sachs and Wu (1977) p. 41 defines an observer as a future-directed timelike curve with a unit tangent vector field]. However, in many cases both of these two factors are critical, and in such cases one should interpret an observer as a world line $G(\tau)$ equipped with a specific orthonormal tetrad field [where $(e_0)^a$ equals the 4-velocity]. This world line describes the orbital motion of the observer (as a point mass), while the spatial angular velocity ω^a of the triad field describes the rotation of the observer. As we mentioned at the end of Sect. 6.3, an inertial observer in Minkowski spacetime refers to a non-rotating ($\omega^a = 0$) observer whose world line is a geodesic (the 4-acceleration $A^a = 0$). This is the simplest type of observer. Similarly, a free-falling ($A^a = 0$) non-rotating ($\omega^a = 0$) observer in curved spacetime also belongs to the simplest type of observer, which has great significance for understanding the equivalence principle and the concept of a local inertial frame (see Sect. 7.5 for details). A clear understanding on the two factors of an observer will be very helpful for distinguishing an inertial force and a Coriolis force (see Sect. 7.4 for details).

[Optional Reading 7.3.1]

To prove Propositions 7.3.5 and 7.3.7, it is necessary to have a quantitative discussion for the gauge freedom of Ω_{ab} . Suppose Ω_{ab} is the angular velocity for the spacetime rotation of a spatial vector field w^a on $G(\tau)$, i.e.,

$$\frac{Dw^a}{d\tau} = -\Omega^{ab} w_b . \quad (7.3.13)$$

Choose an orthonormal tetrad field on $G(\tau)$ such that $(e_0)^a = Z^a$, $(e_1)^a = \alpha w^a$ (where α is the normalization factor), then a necessary and sufficient condition for $\Omega'_{ab} \equiv \Omega_{ab} + \Lambda_{ab}$ to satisfy (7.3.13) is that $\Lambda^{ab}(e^1)_b = 0$. Thus,

$$0 = \Lambda^{ab}(e^1)_b = \Lambda^{\mu\nu}(e_\mu)^a(e_\nu)^b(e^1)_b = \Lambda^{01}(e_0)^a = \Lambda^{01}(e_0)^a + \Lambda^{21}(e_2)^a + \Lambda^{31}(e_3)^a ,$$

and hence $\Lambda_{01} = \Lambda_{21} = \Lambda_{31} = 0$. Since $\Lambda^{ab}(e^1)_b = 0$ is the only restriction on Λ_{ab} , and there is no restriction on the other 3 components Λ_{02} , Λ_{03} and Λ_{23} , one can choose Ω_{02} , Ω_{03} and Ω_{23} arbitrarily. This is the gauge freedom of the spacetime angular velocity Ω_{ab} of w^a .

Proof of Proposition 7.3.5 Choose an orthonormal tetrad field such that $(e_0)^a = Z^a$, $(e_1)^a = \alpha w^a$ (where α is the normalization factor). It follows from

$$0 = \frac{D}{d\tau}(Z^a w_a) = w_a \frac{DZ^a}{d\tau} + Z_a \frac{Dw^a}{d\tau} = -w_a \tilde{\Omega}^{ab} Z_b - Z_a \Omega^{ab} w_b \\ = (\tilde{\Omega}^{ab} - \Omega^{ab}) Z_a w_b = (\Omega^{ab} - \tilde{\Omega}^{ab})(e^0)_a (e^1)_b \alpha^{-1} = (\Omega^{01} - \tilde{\Omega}^{01}) \alpha^{-1}$$

that $\Omega^{01} = \tilde{\Omega}^{01}$. Using the gauge freedom of Ω^{ab} we can let $\Omega^{02} = \tilde{\Omega}^{02}$ and $\Omega^{03} = \tilde{\Omega}^{03}$. Noticing that $\hat{\Omega}_{ij} = 0$, we see that $\hat{\Omega}_{ab} \equiv \Omega_{ab} - \tilde{\Omega}_{ab} = \Omega_{ij}(e^i)_a (e^j)_b$ is a pure spatial rotation. \square

Proof of Proposition 7.3.7 Let $(\hat{\Omega}_i)_{ab}$ represent the angular velocity 2-form for the spatial rotation of $(e_i)^a$. From

$$0 = \frac{D[(e_1)^a (e_2)_a]}{d\tau} = \frac{D[(e_2)^a (e_3)_a]}{d\tau} = \frac{D[(e_3)^a (e_1)_a]}{d\tau}$$

we can see that

$$(a) (\hat{\Omega}_1)^{12} = (\hat{\Omega}_2)^{12}, \quad (b) (\hat{\Omega}_2)^{23} = (\hat{\Omega}_3)^{23}, \quad (c) (\hat{\Omega}_3)^{31} = (\hat{\Omega}_1)^{31}. \quad (7.3.14)$$

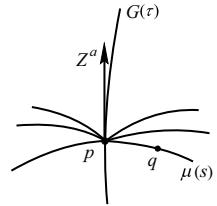
Using the freedom of $(\hat{\Omega}_1)^{23}$ one can set it to equal $(\hat{\Omega}_2)^{23}$, and thus (b) in the equation above can be developed into $(\hat{\Omega}_1)^{23} = (\hat{\Omega}_2)^{23} = (\hat{\Omega}_3)^{23}$. Similarly we have $(\hat{\Omega}_1)^{12} = (\hat{\Omega}_2)^{12} = (\hat{\Omega}_3)^{12}$ and $(\hat{\Omega}_1)^{31} = (\hat{\Omega}_2)^{31} = (\hat{\Omega}_3)^{31}$, and hence $(\hat{\Omega}_1)^{ab} = (\hat{\Omega}_2)^{ab} = (\hat{\Omega}_3)^{ab}$. The reader should prove that $\hat{\Omega}_{ab}$ no longer has any gauge freedom. \square

[The End of Optional Reading 7.3.1]

7.4 The Proper Coordinate System of an Arbitrary Observer

The tetrad of an observer is only defined on the world line of the observer. In order to record the events (experimental results) near the world line, one needs to extend this tetrad in some way and form a coordinate system. We certainly want the coordinate basis of this system on the world line to coincide with the tetrad of the observer. This section will introduce a coordinate system which satisfies this requirement and is quite convenient, called the **proper coordinate system** of an observer. This system should be determined by two ingredients of the observer—the world line $G(\tau)$ and the orthonormal tetrad field on $G(\tau)$. Since we will talk about general observers, $G(\tau)$ is not necessarily a geodesic, and it can have an arbitrary 4-acceleration \hat{A}^a (the hat stands for the 4-acceleration of the observer, as distinguished from the 4-acceleration of a point mass being measured). Also, the orthonormal triad field $\{(e_i)^a\}$ is not necessarily Fermi transported along $G(\tau)$, but can have an arbitrary angular velocity w^a . Of course, both ω^a and \hat{A}^a are spatial vector fields on $G(\tau)$, i.e., $\omega^a Z_a = 0$, $\hat{A}^a Z_a = 0$. Suppose $\mu(s)$ is an arbitrary spacelike geodesic that starts from p on $G(\tau)$ and is orthogonal to $G(\tau)$ at p , where s is the affine parameter that is equal to the arc length, i.e., $T^a \equiv (\partial/\partial s)^a$ is the unit tangent vector. Let q be a point near $G(\tau)$, then there exists a unique spacelike geodesic $\mu(s)$ passing through q . [See Fig. 7.4. If q is far from $G(\tau)$, then there may be more than one such geodesic, or there may not be any such geodesic. Luckily, the observer G only cares about events

Fig. 7.4 Defining the proper coordinates of q relative to G



close to themselves]. Suppose the spacelike geodesic $\mu(s)$ passing through q starts from a point $p = \mu(0)$ on G , we would like to define four coordinates (called **proper coordinates**) t, x^1, x^2, x^3 for q using this geodesic $\mu(s)$. Suppose V_p is the tangent space of p , and W^p is the 3-dimensional subspace in V_p that is orthogonal to $Z^a|_p$, then $T^a|_p \in W^p$. Denote $T^a|_p$ as w^a for short, and denote its components in $(e_i)^a$ as w^i , then the four proper coordinates of q are defined as

$$t(q) := \tau_p, \quad x^i(q) := s_q w^i, \quad i = 1, 2, 3, \quad (7.4.1)$$

where τ_p is the proper time of p (as a point on G), and s_q is the parameter value of $\mu(s)$ at q , namely the arc length of the segment pq on $\mu(s)$. As long as p is near $G(\tau)$, we can use (7.4.1) to define the coordinates, and thus we obtain the proper coordinate system $\{t, x^i\}$ of the observer G , whose coordinate patch is an open neighborhood of $G(\tau)$ [or of a segment of $G(\tau)$]. As the simplest example, we point out that any Lorentzian coordinate system in 4-dimensional Minkowski spacetime can be viewed as the proper coordinate system of the inertial observer whose world line is an x^0 -coordinate line of this system. (Note that the word “inertial” has already required the triad to be Fermi transported along the curve, which is parallelly transported here).

Proposition 7.4.1 *The coordinate basis vectors of a proper coordinate system at any point $p \in G(\tau)$ are identical to the orthonormal tetrad of the observer $G(\tau)$, and therefore the components of a metric $g_{ab}|_p$ in a proper coordinate system are $g_{\mu\nu}|_p = \eta_{\mu\nu}$.*

Proof Let $(e_1)^a$ represent the first basis vector of the orthonormal tetrad at p , and treat it as the w^a we mentioned above. The proper coordinates of each point on the spacelike geodesic $\mu_1(s)$ determined by $(e_1)^a$ satisfy $x^2 = x^3 = 0$, $t = \tau_p$, and thus $\mu_1(s)$ is an x^1 -coordinate line. For this curve, $w^1 = 1$ in $x^1(q) = s_q w^1$, and hence $x^1 = s$ for each point on the curve. Thus, the coordinate basis $(\partial/\partial x^1)^a|_p = (\partial/\partial s)^a|_p = w^a = (e_1)^a$. In a similar manner we have $(\partial/\partial x^2)^a|_p = (e_2)^a$ and $(\partial/\partial x^3)^a|_p = (e_3)^a$. Moreover, it is not difficult to see that $G(\tau)$ is the coordinate line for the proper coordinate t , and $t = \tau$ on this curve, and hence $Z^a|_p = (\partial/\partial t)^a|_p$. This indicates that the proper coordinate basis $\{(\partial/\partial x^\mu)^a\}$ coincides with the orthonormal tetrad $\{Z^a|_p, (e_i)^a|_p\}$. Therefore, the components of $g_{ab}|_p$ in the proper coordinate system are $g_{\mu\nu}|_p = \eta_{\mu\nu}$. \square

$g_{\mu\nu}|_p = \eta_{\mu\nu}$ is a major feature of the proper coordinate system. Of course, this simple result does not necessarily hold for a point outside $G(\tau)$.

A proper coordinate system has many uses. For example, by means of it one can define the 3-velocity and 3-acceleration for a point mass.

Definition 1 Suppose $\{t, x^i\}$ is the proper coordinate system of an observer G , and (at least a segment L of) the world line of a point mass is located in the proper coordinate patch of G , then the **3-velocity** u^a and the **3-acceleration** a^a are accordingly defined as

$$u^a := \frac{dx^i(t)}{dt} \left(\frac{\partial}{\partial x^i} \right)^a, \quad (7.4.2)$$

$$a^a := \frac{d^2x^i(t)}{dt^2} \left(\frac{\partial}{\partial x^i} \right)^a, \quad (7.4.3)$$

where $x^i(t)$ are the parametric representations for L with t as the parameter in the proper coordinate system.

Remark 1 If p is the intersection of L and G , then following (6.3.28) we can also define the 3-velocity of L at p relative to the observer G as

$$u^a := \frac{h^a_b U^b}{\gamma}, \quad (7.4.4)$$

where U^a is the 4-velocity of L , $\gamma \equiv -Z^a U_a$, $h_{ab} \equiv g_{ab} + Z_a Z_b$, and $h^a_b \equiv g^{ac} h_{cb}$. Now we will show that (7.4.4) is equivalent to (7.4.2). Suppose τ_L is the proper time of the point mass L , then the 4-velocity $U^a = (\partial/\partial\tau_L)^a$ at p can be expanded in terms of the proper coordinate basis as

$$U^a = \left(\frac{\partial}{\partial t} \right)^a \frac{dt}{d\tau_L} + \left(\frac{\partial}{\partial x^i} \right)^a \frac{dx^i}{d\tau_L}.$$

In the equation above, $(\partial/\partial t)^a$ is exactly Z^a , whose spatial projection vanishes; $(\partial/\partial x^i)^a$ is orthogonal to Z^a , and thus its projection is equal to itself. Hence,

$$h^a_b U^b = \left(\frac{\partial}{\partial x^i} \right)^a \frac{dx^i}{d\tau_L}. \quad (7.4.5)$$

Using the proper coordinate system, one can also find another expression for $\gamma \equiv -Z^a U_a$:

$$\begin{aligned} \gamma &= -g_{ab} Z^a U^b|_p = -g_{\mu\nu} Z^\mu U^\nu|_p = -\eta_{\mu\nu} (\partial/\partial t)^\mu U^\nu|_p, \\ &= -\eta_{00} (\partial/\partial t)^0 U^0|_p = U^0|_p = dt/d\tau_L|_p, \end{aligned} \quad (7.4.6)$$

where Proposition 7.4.1 is used in the third equality. It follows from (7.4.5) and (7.4.6) that $h^a{}_b U^b / \gamma = (\partial / \partial x^i)^a dx^i / dt$, and thus (7.4.4) is equivalent to (7.4.2).

The 3-velocity defined above can help deepen the understanding of inertial forces and Coriolis forces in Newtonian mechanics (and their generalizations in curved spacetime). According to Newtonian mechanics, Newton's second law does not hold when a non-inertial observer G measures the motion of a point mass. To preserve the form of this law, people introduced the concept of a fictitious force. Suppose the 3-acceleration of G relative to an inertial frame is $\hat{\vec{a}}$. (The hat is added to represent the 3-acceleration of the *observer*, in order to distinguish from the 3-acceleration \vec{a} of the point mass being measured). When G makes a measurement, if they regard any point mass L being measured as experiencing an imaginary inertial force $-m\hat{\vec{a}}$ (where m is the mass of the point mass), then the equation of motion of a free point mass after the inertial force is taken into account is $-m\hat{\vec{a}} = m\vec{a}$, and thus the 3-acceleration of L relative to G is $\vec{a} = -\hat{\vec{a}}$. This can be called the inertial acceleration of L relative to G which, when multiplied by m , is the inertial force. (We stipulate that the observer and the world line of the point mass intersect, and the measurement is made at the intersection). When G is rotating, however, a Coriolis force must be introduced in addition to the inertial force to preserve the form of Newton's second law. However, the phrase “the observer is rotating” may sometimes cause confusion, so it is necessary to discuss this in greater detail.

Consider a large rigid disk which rotates around its own axis. A swivel chair is put on the edge of the disk, and the chair base is fixed on the disk (but the chair can rotate around the axis fixed on its base). Due to the rotation of the disk, the observer in the swivel chair undergoes a circular motion (the world line is a helix), which is a special case of orbital motion. Of course, the observer in the swivel chair can also rotate with respect to their own axis. (This motion is unrelated to the shape of the world line; it is described by the motion of the orthonormal frame attached to the observer along the world line). Since the observer has been regarded as a point mass, and the motion of a point mass cannot be separated as a rotation and a translation, “the orbital motion of an observer on a rotating disk is circular motion” is the most accurate way to refer to this type of motion. However, in our daily life we also often refer to the circular motion of a point mass as a rotation, which can be easily confused with the rotation of its frame. Unfortunately, distinguishing orbital motion and a frame rotation happens to be the key for distinguishing inertial forces and Coriolis forces. Therefore, we refer to the circular motion (a special case of orbital motion) of the observer caused by the rotation of the disk and the rotation of the frame realized using a swivel chair as **revolution** and **rotation**, respectively. This is similar to calling the Earth's (viewed as a point mass) circular motion around the Sun as revolution, while calling the Earth's (now treated as a rigid body) rotation around its axis as rotation. Certainly, the word revolution is not as appropriate as the term orbital motion when the world line of the observer is not a helix. Later we will see that inertial forces and Coriolis forces originate from the orbital motion and the rotation of the observer, *respectively*. Now let us have a quantitative discussion with an arbitrary spacetime as the background; in the low speed approximation, the conclusions for Minkowski spacetime agree

with Newtonian mechanics. For simplicity, we only discuss the measurements on a *free point mass* by an arbitrary observer G . Although L is a free point mass, the arbitrariness of the observer G (including the fact that the world line may not be a geodesic and the orthonormal spatial triad may not be Fermi transported) means that measurements of L by G will have an inertial acceleration and a Coriolis acceleration. [An inertial (Coriolis) force in Newtonian mechanics is equal to the inertial (Coriolis) acceleration times the mass of L]. See the following proposition.

Proposition 7.4.2 *Suppose an observer G has a 4-acceleration \hat{A}^a and an angular velocity ω^a (i.e., the angular velocity for the rotation of its spatial triad). Also suppose the world lines of G and the free point mass L being measured intersect at p , and the 3-velocity of L at p relative to G is u^a . Then the 3-acceleration of L at p relative to G is*

$$a^a \equiv (d^2x^i/dt^2)(e_i)^a = -\hat{A}^a - 2\varepsilon_{bc}^a \omega^b u^c + 2(\hat{A}_b u^b)u^a, \quad (7.4.7)$$

where $(e_i)^a$ is the orthonormal spatial triad of the observer at p , $\varepsilon_{abc} \equiv Z^d \varepsilon_{dabc}$, Z^d is the 4-velocity of G at p , and ε_{abcd} is the volume element associated with the spacetime metric g_{ab} .

Proof See Optional Reading 7.4.1. □

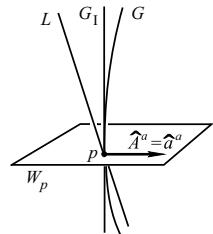
Now let us discuss the physical meaning of each term on the right-hand side of (7.4.7). If G is a freely falling non-rotating observer (for Minkowski spacetime this is an inertial observer), i.e., $\hat{A}^a = 0$, $\omega^a = 0$, then from (7.4.7) we can see that the 3-acceleration of L as measured by G is $a^a = 0$. Take Minkowski spacetime as an example, this indicates nothing but the simple fact that there is only a relative velocity but no relative acceleration between two point masses undergoing inertial motion. In contrast, if G is not a freely falling non-rotating observer, then there are the following three possibilities:

(a) The world line of G is not a geodesic ($\hat{A}^a \neq 0$), but G is still a non-rotating observer ($\omega^a = 0$, i.e., its tetrad is Fermi transported along the world line). Now (7.4.7) becomes

$$a^a = -\hat{A}^a + 2(\hat{A}_b u^b)u^a. \quad (7.4.8)$$

Let \hat{A} and u represent the magnitudes of the spatial vectors \hat{A}^a and u^a , and let θ be the angle between them. Then the magnitude of the second term on the right-hand side of the above equation is $2\hat{A}u^2 \cos \theta \leq 2\hat{A}u^2$, and hence the second term can be neglected under the non-relativistic approximation $u \ll 1$. For Minkowski spacetime, suppose G_I is the instantaneous rest inertial observer of G at p (see Fig. 7.5), and \hat{a}^a is the 3-acceleration of G relative to G_I , then it follows from Proposition 6.3.6 that $\hat{a}^a = \hat{A}^a$. Since in Newtonian mechanics, $-\hat{a}^a$ is exactly the inertial acceleration added for a point mass when observed by a non-inertial observer, the first term $-\hat{A}^a$ on the right-hand side of (7.4.8) can be interpreted as an inertial acceleration, and the second term is the relativistic correction term for the inertial acceleration (which vanishes under the Newtonian approximation $u \ll 1$). For curved spacetime, it can be

Fig. 7.5 The 3-acceleration \hat{a} of G relative to the instantaneous rest inertial observer G_I is equal to the 4-acceleration \hat{A}^a of G



proved that (Lemma 7.4.3 is used, left as Exercise 7.6) as long as we interpret G_I as the freely falling observer that is at rest relative to G at p , then we still have $\hat{a}^a = \hat{A}^a$ (\hat{a}^a is the 3-acceleration of G relative to G_I), and hence the first and second terms on the right-hand side of (7.4.8) can still be interpreted as the inertial acceleration and the corresponding correction term, respectively. In conclusion, the inertial acceleration is caused by the 4-acceleration \hat{A}^a of the observer (which depends on its orbital motion).

(b) The world line of G is a geodesic ($\hat{A}^a = 0$), but G has a rotation ($\omega^a \neq 0$), such as a rotating observer in the swivel chair fixed on the floor of a freely falling spaceship. Now (7.4.7) becomes

$$a^a = -2\varepsilon^a_{bc}\omega^b u^c = 2\vec{u} \times \vec{\omega}. \quad (7.4.9)$$

This 3-acceleration of the free point mass L relative to G comes completely from the rotation of the observer ($\omega^a \neq 0$). The right-hand side of the equation above is the same as the expression for the Coriolis acceleration in Newtonian mechanics, and hence in curved spacetime is also called the Coriolis acceleration. This clearly indicates the difference between an inertial acceleration and a Coriolis acceleration: the former originates from the non-geodesic motion of the observer, while the latter comes from the rotation of the observer. In the case of a rotating disk, many textbooks on mechanics assume that the observer on the rotating disk must have a corresponding rotation due to the revolution, and attribute Coriolis forces to the revolution of the observer. Actually, the rotation and revolution of the observer on a disk are in principle independent. Suppose an observer is holding a gyroscope, sitting in a swivel chair whose base is fixed on the edge of the disk. Then the observer can adjust (“rotate”) the swivel chair properly and always face the direction indicated by the gyroscope, and thus is non-rotating while revolving with the disk. In this case, a point mass being measured will only have an inertial acceleration but no Coriolis acceleration!

(c) The world line of G is not a geodesic ($\hat{A}^a \neq 0$), and G has a rotation ($\omega^a \neq 0$). A free point mass observed by G will have both an inertial acceleration and a Coriolis acceleration.

Many authors regard Coriolis force as a type of inertial force, this is nothing but a problem of name, which is totally fine. However, in order to distinguish the orbital motion and rotation of an observer, this text prefers the name used by some other authors [e.g., Misner et al. (1973)], i.e., to call the fictitious forces caused by

the orbital motion and rotation of an observer as inertial forces and Coriolis forces, respectively.

[Optional Reading 7.4.1]

To prove Proposition 7.4.2, we first prove the following Lemma.

Lemma 7.4.3 *The Christoffel symbols of the spacetime metric g_{ab} in the proper coordinate system of $G(\tau)$ have the following simple forms:*

$$\begin{aligned}\Gamma^0_{00} &= \Gamma^\sigma_{ij} = 0, & \Gamma^0_{0i} &= \Gamma^0_{i0} = \Gamma^i_{00} = \hat{A}_i, \\ \Gamma^i_{0j} &= \Gamma^i_{j0} = -\omega^k \varepsilon_{0kij}, & \sigma &= 0, 1, 2, 3, \quad i, j, k = 1, 2, 3.\end{aligned}\tag{7.4.10}$$

where \hat{A}^a and ω^a are the 4-acceleration and spatial angular velocity of the observer G , respectively, and ε_{0kij} are the components of the volume element associated with g_{ab} in the proper coordinate system.

Proof Since the orthonormal triad $\{(e_i)^a\}$ of the observer G has a spatial rotation with an angular velocity ω^a , from Sect. 7.3 we know that

$$(e_0)^b \nabla_b (e_\mu)^a = D(e_\mu)^a / d\tau = -\Omega^{ab} (e_\mu)_b, \quad \mu = 0, 1, 2, 3,\tag{7.4.11}$$

where

$$\Omega_{ab} = \hat{A}_a \wedge Z_b + \varepsilon_{abc} \omega^c.\tag{7.4.12}$$

From (5.7.2) we know the Christoffel symbols satisfy the following equation:

$$(\partial/\partial x^\nu)^b \nabla_b (\partial/\partial x^\mu)^a = \Gamma^\sigma_{\mu\nu} (\partial/\partial x^\sigma)^a,\tag{7.4.13}$$

where $\{(\partial/\partial x^\mu)^a\}$ are the coordinate basis in the coordinate system associated with the Christoffel symbols. Now we are in the proper coordinate system of $G(\tau)$, and the proper coordinate basis is the same as the orthonormal frame. Hence, (7.4.13) on $G(\tau)$ can also be expressed as

$$(e_0)^b \nabla_b (e_\mu)^a = \Gamma^\sigma_{\mu 0} (e_\sigma)^a.\tag{7.4.14}$$

Comparing (7.4.11) and (7.4.14) yields $\Gamma^\sigma_{\mu 0} (e_\sigma)^a = -\Omega^a{}_b (e_\mu)^b = -\Omega^\sigma{}_\mu (e_\sigma)^a$, and therefore

$$\Gamma^\sigma_{\mu 0} = -\Omega^\sigma{}_\mu, \quad \sigma, \mu = 0, 1, 2, 3.$$

If we rewrite (7.4.14) into the component form, then the above equation becomes

$$\Gamma^\sigma_{\mu 0} = -(\hat{A}^\sigma Z_\mu - Z^\sigma \hat{A}_\mu + Z_\alpha \omega_\rho \varepsilon^{\alpha\rho\sigma}{}_\mu) = -(\hat{A}^\sigma Z_\mu - Z^\sigma \hat{A}_\mu + \omega_\rho \varepsilon^{0\rho\sigma}{}_\mu),$$

where in the last step we used $Z_i = 0$ and $Z_0 = -1$. Using also $Z^0 = 1$, $\hat{A}^0 = 0 = \hat{A}_0$, we have

$$\begin{aligned}\Gamma^0_{00} &= -(\hat{A}^0 Z_0 - Z^0 \hat{A}_0 - \omega_\rho \varepsilon^{0\rho 0}{}_0) = 0, \\ \Gamma^0_{i0} &= -(\hat{A}^0 Z_i - Z^0 \hat{A}_i - \omega_\rho \varepsilon^{0\rho 0}{}_i) = \hat{A}_i, \\ \Gamma^i_{00} &= -(\hat{A}^i Z_0 - Z^i \hat{A}_0 - \omega_\rho \varepsilon^{0\rho i}{}_0) = \hat{A}_i, \\ \Gamma^i_{j0} &= -(\hat{A}^i Z_j - Z^i \hat{A}_j - \omega_\rho \varepsilon^{0\rho i}{}_j) = \omega_k \varepsilon^{0ki}{}_j = -\omega^k \varepsilon_{0kij},\end{aligned}$$

where $i, j, k = 1, 2, 3$. Finally, we show that $\Gamma^\sigma{}_{ij} = 0$. Suppose $\mu(s)$ is a spacelike geodesic starting from $p \in G$ (where s is the arc length), whose tangent vector T^a at p is orthogonal to Z^a , then along $\mu(s)$ we have

$$x^0 \equiv t = \tau_p = \text{constant}, \quad x^i = s T^i, \quad T^i = \text{constant}, \quad i = 1, 2, 3.$$

Thus, $d^2x^\sigma/ds^2 = 0$, $\sigma = 0, 1, 2, 3$. Hence, from the geodesic equation we have

$$0 = \frac{d^2x^\sigma}{ds^2} + \Gamma^\sigma_{\mu\nu} \frac{dx^\mu}{ds} \frac{dx^\nu}{ds} = \Gamma^\sigma_{ij} \frac{dx^i}{ds} \frac{dx^j}{ds}, \quad \sigma = 0, 1, 2, 3.$$

That is, $0 = \Gamma^\sigma_{ij} T^i T^j$ ($i = 1, 2, 3$) \forall unit vectors $T^a \in W_p$, and thus $0 = \Gamma^\sigma_{ij} w^i w^j$, $\forall w^a \in W_p$. Therefore at p we have $\Gamma^\sigma_{ij} = 0$, $i, j = 1, 2, 3$ and $\sigma = 0, 1, 2, 3$. Since $p \in G$ is arbitrary, this equation holds for any point on $G(\tau)$. \square

Proof of Proposition 7.4.2 The world line of a free point mass is a geodesic, and its equation in the proper coordinate system of $G(\tau)$ is

$$\frac{d^2x^\mu}{d\tau_L^2} + \Gamma^\mu_{\nu\sigma} \frac{dx^\nu}{d\tau_L} \frac{dx^\sigma}{d\tau_L} = 0, \quad (7.4.15)$$

where the affine parameter τ_L of the geodesic is the proper time of the point mass L . Choose $t \equiv x^0$ as another parameter [the coordinate time of the proper coordinate system of $G(\tau)$] of L , and denote $dt/d\tau_L$ as γ . Then, $\frac{dx^\mu}{d\tau_L} = \frac{dx^\mu}{dt} \frac{dt}{d\tau_L} = \gamma \frac{dx^\mu}{dt}$, and hence

$$\frac{d^2x^\mu}{d\tau_L^2} = \gamma \frac{d}{dt} \left(\frac{dx^\mu}{d\tau_L} \right) = \gamma \frac{d}{dt} \left(\gamma \frac{dx^\mu}{dt} \right) = \gamma \left(\gamma \frac{d^2x^\mu}{dt^2} + \frac{d\gamma}{dt} \frac{dx^\mu}{dt} \right). \quad (7.4.16)$$

Setting $\mu = i$ ($= 1, 2, 3$) in the above equation yields

$$\frac{d^2x^i}{d\tau_L^2} = \gamma \left(\gamma a^i + \frac{d\gamma}{dt} u^i \right). \quad (7.4.17)$$

Setting $\mu = i$ in (7.4.15), and plugging in (7.4.17), we get

$$\gamma \left(\gamma a^i + \frac{d\gamma}{dt} u^i \right) + \Gamma^i_{\nu\sigma} \frac{dx^\nu}{dt} \frac{dx^\sigma}{dt} \gamma^2 = 0.$$

Hence,

$$\begin{aligned} a^i &= -\gamma^{-1} u^i d\gamma/dt - (\Gamma^i_{00} + 2\Gamma^i_{0j} u^j + \Gamma^i_{jk} u^j u^k) \\ &= \gamma^{-1} u^i d\gamma/dt - (\hat{A}^i - 2\omega^k \varepsilon_{0kij} u^j) = -\gamma^{-1} u^i d\gamma/dt - \hat{A}^i - 2\varepsilon^i_{jk} \omega^j u^k, \end{aligned} \quad (7.4.18)$$

where in the second equality we used Lemma 7.4.3, and in the third equality we used $\varepsilon_{0kij} = \varepsilon_{kij}$. To derive $\gamma^{-1} d\gamma/dt$, we set $\mu = 0$ in (7.4.16), and find $d^2t/d\tau_L^2 = \gamma dt/d\tau_L$. Then setting $\mu = 0$ in (7.4.15) yields

$$0 = \frac{d^2t}{d\tau_L^2} + \Gamma^0_{\nu\sigma} \frac{dx^\nu}{d\tau_L} \frac{dx^\sigma}{d\tau_L} = \gamma \frac{d\gamma}{dt} + 2\Gamma^0_{0i} \frac{dt}{dt} \frac{dx^i}{dt} \gamma^2 = \gamma \frac{d\gamma}{dt} + 2\hat{A}_i u^i \gamma^2,$$

where Lemma 7.4.3 is used in both the second and third equalities. From the above equation we get $-\gamma^{-1} \frac{d\gamma}{dt} = 2\hat{A}_b u^b$. Plugging this into (7.4.18) and rewriting it using the abstract indices, we obtain $a^a = -\hat{A}^a - 2\varepsilon^a_{bc} \omega^b u^c + 2(\hat{A}_b u^b) u^a$. \square

[The End of Optional Reading 7.4.1]

7.5 Equivalence Principles and Local Inertial Frames

Any inertial coordinate system in Minkowski spacetime is globally defined (the coordinate patch covers the whole manifold), and thus is also called a global inertial coordinate system. Suppose $\{t, x, y, z\}$ is a global inertial coordinate system, and $G(\tau)$ is an arbitrary t -coordinate line in this system, then $\{(\partial/\partial t)^a, (\partial/\partial x)^a, (\partial/\partial y)^a, (\partial/\partial z)^a\}|_p$ is a non-rotating orthonormal tetrad field on G . This coordinate line together with this tetrad field form an inertial observer, and $\{t, x, y, z\}$ is exactly the proper coordinate system of this observer. Now we discuss to what extent can the concepts above be generalized to curved spacetime. First, an inertial observer in Minkowski spacetime corresponds naturally to a freely falling (the world line is a geodesic) non-rotating observer. To study the results of a measurement from this observer, let us discuss the famous “Einstein’s elevator”. Suppose an elevator close to the ground is freely falling due to the breaking of the cable. The rest observer inside this elevator will experience weightlessness, which is already a fact in Newtonian mechanics. Suppose he lets go of an apple in his hand, he will find that the apple does not fall as usual but instead remains at rest. The reason is very simple: the elevator observer G has a gravitational acceleration \vec{g} relative to an inertial frame (the Earth), and thus is a non-inertial observer (based on the viewpoint of Newtonian mechanics, based on general relativity it will be the opposite). Hence, he will think there are two forces being applied to the apple: the gravitational force $m_G \vec{g}$ (where m_G is the gravitational mass of the apple) and the inertial force $-m_I \vec{g}$ (where m_I is the inertial mass of the apple). Since $m_G = m_I$ (which is the crux of this argument), the net force vanishes, and therefore it is unaccelerated, or, in a state of weightlessness. If he is an astronaut, he will feel that this apple behaves the same as an apple in an inertial spaceship far away from celestial bodies (and thus the spacetime is approximately flat). By extension, since $m_G = m_I$, according to Newtonian mechanics, every (non-gravitational)⁷ mechanical experiment in Einstein’s elevator has the same result as the corresponding experiment in an inertial spaceship far way from celestial bodies. This is exactly the reason why $m_G = m_I$ is called an equivalence principle.

During the process of conceiving general relativity, Einstein carried over this principle further hypothetically from mechanical experiments to all physical experiments, i.e., he assumed every (non-gravitational) experiment in a freely falling elevator has the same results as the corresponding experiment in an inertial spaceship far away from celestial bodies (i.e., in flat spacetime). Based on this, he derived conclusions like the gravitational redshift of light and light rays following curved paths in a gravitational field. The principle corresponding to $m_G = m_I$ is dubbed the **weak equivalence principle** (WEP), and the one generalized by Einstein is dubbed the **Einstein equivalence principle** (EEP). Now we will discuss this principle from the perspective of general relativity.

⁷ A non-gravitational experiment refers to an experiment where the gravitational interaction between the objects in the lab can be ignored, but there may exist a gravitational field produced by an object outside the lab (e.g., the Earth).

Proposition 7.5.1 Suppose $G(\tau)$ is a freely falling non-rotating observer in curved spacetime (e.g., an observer in Einstein's elevator), $g_{\mu\nu}$ are the components of the metric g_{ab} in the proper coordinate system $G(\tau)$, and $\Gamma^\sigma_{\mu\nu}$ are the Christoffel symbols of the derivative operator ∇_a associated with g_{ab} in this system, then

$$g_{\mu\nu}|_p = \eta_{\mu\nu}, \quad \Gamma^\sigma_{\mu\nu}|_p = 0 \quad (\sigma, \mu, \nu = 0, 1, 2, 3), \quad \forall p \in G. \quad (7.5.1)$$

Proof $g_{\mu\nu}|_p = \eta_{\mu\nu}$ is the conclusion of Proposition 7.4.1 (which holds for the proper coordinate system of any observer). Lemma 7.4.3 gives $\Gamma^\sigma_{\mu\nu}|_p = 0$ ($\sigma, \mu, \nu = 0, 1, 2, 3$) when $G(\tau)$ is a geodesic and the corresponding observer is non-rotating. \square

Taking the electromagnetic phenomenon as an example, let us discuss the applications of the above proposition. According to the minimal substitution rule (see Sect. 7.2), the expressions for Maxwell's equations and the equation of the Lorentz force in curved spacetime are

$$(a) \nabla^a F_{ab} = -4\pi J_b, \quad (b) \nabla_{[a} F_{bc]} = 0, \quad (c) q F^a{}_b U^b = U^b \nabla_b P^a \equiv \frac{DP^a}{d\tau}. \quad (7.5.2)$$

Suppose $\{x^\mu\}$ is an arbitrary local coordinate system, we want to write down the expressions for the components of (7.5.2) in this system. First we look at (a). Recall that the coordinate components of $\nabla_a v^b$ are denoted by $v^\nu{}_{;\mu}$ (see Sect. 3.1), i.e., $v^\nu{}_{;\mu} \equiv (dx^\nu)_b (\partial/\partial x^\mu)^a \nabla_a v^b$. Similarly, one should denote the coordinate components of $\nabla_a F^c{}_b$ as $F^\sigma{}_{v;\mu}$, i.e., $F^\sigma{}_{v;\mu} \equiv (dx^\sigma)_c (\partial/\partial x^\mu)^a (\partial/\partial x^\nu)^b \nabla_a F^c{}_b$. Hence,

$$F^\mu{}_{v;\mu} = (dx^\mu)_c \left(\frac{\partial}{\partial x^\mu} \right)^a \left(\frac{\partial}{\partial x^\nu} \right)^b \nabla_a F^c{}_b = \delta^a{}_c \left(\frac{\partial}{\partial x^\nu} \right)^b \nabla_a F^c{}_b = \left(\frac{\partial}{\partial x^\nu} \right)^b \nabla_a F^a{}_b$$

are the coordinate components of $\nabla_a F^a{}_b$, and the component expression for (7.5.2)(a) is

$$F^\mu{}_{v;\mu} = -4\pi J_v. \quad (7.5.3a)$$

Similarly, the coordinate components of $\nabla_a F_{bc}$ are denoted by $F_{v\sigma;\mu}$, and the coordinate component expression for (7.5.2)(b) is

$$F_{[v\sigma;\mu]} = 0. \quad (7.5.3b)$$

Finally, for (7.5.2)(c), the coordinate components of the left-hand side are obviously $q F^\mu{}_\nu U^\nu$. Using $DP^\mu/d\tau$ to represent the coordinate components of $DP^a/d\tau$, we have

$$q F^\mu{}_\nu U^\nu = \frac{DP^\mu}{d\tau}. \quad (7.5.3c)$$

Note that in general $D P^a / d\tau \neq dP^a / d\tau$, because it is not difficult to show that (see Exercise 3.6) $D P^a / d\tau = dP^a / d\tau + \Gamma^a_{v\sigma} U^v P^\sigma$. Since $\forall p \in G$ we have $\Gamma^\mu_{v\sigma}|_p = 0$ for the proper coordinate system of G , the above equations can be written as

$$(a) F^\mu_{v;\mu} = -4\pi J_v, \quad (b) F_{[v\sigma;\mu]} = 0, \quad (c) q F^\mu_v U^\nu = \frac{dP^\mu}{d\tau}. \quad (7.5.4)$$

These are exactly the expressions for the corresponding laws in (7.5.2) in a global inertial (Lorentzian) coordinate system in Minkowski spacetime. The discussion above can be generalized to other physical laws. Thus, the proper coordinate system of a freely falling non-rotating observer is similar to a global inertial (Lorentzian) coordinate system, and therefore is called a **local inertial frame**, also called a **local Lorentz system** or **local Lorentz frame**.

People often say: The laws of physics are the same in any local Lorentz system of curved spacetime as in an inertial coordinate system in Minkowski spacetime [Misner et al. (1973) p. 207], and thus all the physical experiments done by a freely falling non-rotating observer G have the same (equivalent) results as the corresponding experiments done by an inertial observer in flat spacetime. This is the conclusion required by the Einstein equivalence principle. However, the statement above is not quite precise, since all we can be certain about is $\Gamma^\sigma_{\mu\nu}|_p = 0, \forall p \in G$, and once one deviates from $G(\tau)$, we cannot guarantee that $\Gamma^\sigma_{\mu\nu} = 0$. In fact, if $\Gamma^\sigma_{\mu\nu}$ really vanish in a neighborhood of $G(\tau)$, then $\forall p \in G$ we have

$$R_{\mu\nu\rho}{}^\sigma|_p = (-2\partial_{[\mu}\Gamma^\sigma_{\nu]\rho} + 2\Gamma^\lambda_{\rho[\mu}\Gamma^\sigma_{\nu]\lambda})|_p = 0,$$

i.e., the curvature at each point on $G(\tau)$ vanishes, which is inconsistent with the curved spacetime we supposed. The heart of the problem is that, by choosing a coordinate system one can only make the $\Gamma^\sigma_{\mu\nu}$ on $G(\tau)$ vanish but not the curvature (curvature is independent of the coordinate system). Thus, the statement “the laws of physics are the same in any local inertial frame of curved spacetime as in an inertial frame in Minkowski spacetime” is not necessarily true for a point in the coordinate patch but outside the curve $G(\tau)$. Nevertheless, when the observer G is doing an experiment, a “finitely small” spacetime neighborhood U of the world line is usually involved (e.g., an elevator is involved for an observer in the elevator, see Fig. 7.6), and thus the problem becomes not that simple. Luckily, the effect of spacetime curvature can only be made manifest (detected by experiments) in a sufficiently large spacetime region. Hence, as long as the spacetime neighborhood that is involved in an experiment is sufficiently small (for an elevator, as long as its spatial scale and the falling time are sufficiently small), the result of the experiment will be virtually indistinguishable from the corresponding experiment in flat spacetime. [This is similar to the following simple example: at each point on a 2-dimensional sphere, $R_{abc}{}^d$ is nonvanishing; however, if one only cares about a small piece of the sphere ΔS in the vicinity of a point, then ΔS can be substituted approximately by a small region $\overline{\Delta S}$ of the tangent plane of this point (see Fig. 7.7). For instance, in order to measure the angle between two meridians of the Earth at the North Pole, one may

Fig. 7.6 The “small” spacetime neighborhood U involved in the experiment done by an observer

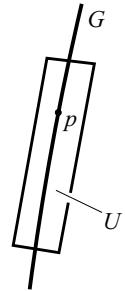
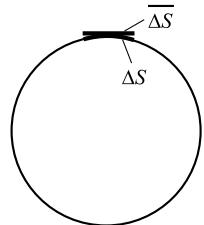


Fig. 7.7 A small piece of sphere ΔS in the vicinity of the north pole can be substituted approximately by a small region $\Delta \bar{S}$ of the tangent plane



treat a small segment of each meridians as a straight line]. In an arbitrary curved spacetime, if one only cares about a point’s neighborhood that is sufficiently small, special relativistic laws of physics can be used as an approximation.

The proper coordinate system of a freely falling non-rotating observer is a coordinate system in curved spacetime that is very similar to an inertial coordinate system in Minkowski spacetime, and thus is called a local Lorentz system. In addition, in curved spacetime, as long as the Christoffel symbols of a coordinate system vanish at a point p in the coordinate patch (i.e., $\Gamma^\sigma_{\mu\nu}|_p = 0$), this system can also be called a **local Lorentz system at p** .

In Newtonian mechanics, suppose a spaceship far away from celestial bodies is undergoing uniformly accelerated motion. An observer in the spaceship (which is an accelerating observer) will see that the apple flying out of their hand undergoes a uniformly accelerated motion in the opposite direction, just like what will happen near the Earth. It is not difficult to believe that all the (non-gravitational) mechanical experiments in the spaceship will have approximately the same results as the corresponding experiments near the ground; this can be regarded as another formulation of the weak equivalence principle. Based on this, people also often say that “an astronaut in an accelerating spaceship finds themselves in a gravitational field” or “an acceleration is equivalent to a gravitational field”. One should have a proper interpretation for these two statements. Based on the first statement, beginners may raise a question like this: since the astronaut in an accelerating spaceship feels gravity, and gravity is the spacetime curvature, does the astronaut not feel that he is in a curved spacetime? The answer is negative: since we have already stipulated that the spaceship is far away from celestial bodies, the spacetime local to where the spaceship is located must be approximately flat, no matter to which observer. The

reason that leads to the incorrect conclusion above is that the word “gravity” is used twice in the deduction, while they have different meanings. The “gravity” felt by the astronaut is only a fictitious apparent gravity, which is not produced by matter and does not correspond to curved spacetime; the name only comes from the feeling of the astronaut.

Physicists have varied opinions about the meaning and the value of the equivalence principles. Especially, the opinions on their value are also different due to different opinions on the meaning of equivalence principles. Some consider that they are of great significance. For example, Misner et al. (1973) p. 386 said that “The principle of equivalence has great power. With it one can generalize all the special relativistic laws of physics to curved spacetime.” They also said (p. 207) “The vehicle that carries one from classical mechanics to quantum mechanics is the correspondence principle. Similarly, the vehicle between flat spacetime and curved spacetime is the equivalence principle.” Some others, however, take a completely opposite view. For example, J. L. Synge wrote in the preface of Synge (1960) that: “I have never been able to understand this principle. Does it mean that the effects of a gravitational field are indistinguishable from the effects of an observer’s acceleration? If so, it is false. In Einstein’s theory, either there is a gravitational field or there is none, according as the Riemann tensor does not or does vanish. This is an absolute property; it has nothing to do with any observer’s world line. Spacetime is either flat or curved, and in several places in this book I have been at considerable pains to separate truly gravitational effects due to curvature of spacetime from those due to curvature of the observer’s world line (in most ordinary cases the latter predominate). The principle of equivalence performed the essential office of midwife at the birth of general relativity, I suggest that the midwife be now buried with appropriate honors and the facts of absolute spacetime be faced.” This view on equivalence principles might be somewhat extreme, but some statements in the quotation above can yet be regarded as a sobering pill which prevents us from misconstruing concepts. For instance, his warning on distinguishing the real gravity caused by the spacetime curvature from the apparent (fake) gravity caused by the observer’s world line being curved (non-geodesic) is extremely necessary.

Here, we talk briefly about our humble understanding of equivalence principles.

Firstly, the Einstein equivalence principle is a hypothetical generalization of the weak equivalence principle posed by Einstein during the conception of general relativity, which is very important as a midwife at the birth of general relativity. Even Synge agreed with this.

Secondly, as mentioned in Sect. 7.2, physical laws in curved spacetime must obey two principles: (a) the principle of general covariance, and (b) when g_{ab} equals η_{ab} , they can go back to the corresponding laws in special relativity. This is the how this text and some other textbooks state them. More textbooks, however, state principle (b) in another way: (b') the Einstein equivalence principle. From (a) and (b') one can obtain their minimal substitution rule: “the equation of a physical law in a local Lorentz system of curved spacetime can be obtained by changing the commas in the equation of the corresponding physical law in a Lorentzian coordinate system of Minkowski spacetime to semicolons (i.e., changing partial derivatives to

covariant derivatives).” Thus, using the Einstein equivalence principle (together with the principle of general covariance) we can obtain the laws of physics in general relativity from the corresponding laws of physics in special relativity, and therefore it can be said to be the “bridge that brings us from spacial relativity to general relativity”. However, just like what we did in Sect. 7.2, one can also get the physical laws of curved spacetime not by mentioning equivalence principles but by saying (in adding to the principle of general covariance) that “the physical laws should go back to the corresponding laws in special relativity when g_{ab} equals η_{ab} ”. (Either way, we obtain the minimal substitution rule).⁸ Once the physical laws in curved spacetime are accepted (and thus general relativity is formulated), one can totally discuss physics problems without using equivalence principles (although many authors like to use equivalence principles in many problems). Therefore, from this perspective, “burying the midwife” seems have no influence on general relativity.

Thirdly, for some complicated situation (such as when talking about if a charged particle moving along a geodesic in curved spacetime has electromagnetic radiation), “whether or not the principle of equivalence is violated” has been a controversial issue for a long time. We think the point is that the precise meaning of the “principle of equivalence” in these situations has yet to be clarified (another important problem is the definition of radiation). In this sense, maybe it is not excessive at all when Synge said “I have never been able to understand this principle.”

Fourthly, besides general relativity, there exist tons of different gravitational theories [see, for example, Will (2018)]. All the gravitational theories can be classified into two major kinds, namely metric theories (which require the spacetime to have a metric, and the world line of a free point mass is the geodesic of this metric, etc.) and non-metric theories. General relativity is of course a metric theory. There are also many other metric theories out there. For example, another famous and competitive metric theory is called the Brans-Dicke theory, in which the quantities describing gravity also contains a scalar field ϕ other than the metric field g_{ab} . The criterion for judging which gravitational theory is the correct one is of course experiments. To this end, we need a theory about gravitational experiments. R. H. Dicke had been working on this kind of theory since the 1960s. His pioneering works have gradually deepened people’s understanding of equivalence principles and their meaning. At last, people realized that one should put equivalence principles at the important position of inspecting the foundation of gravitational theories (not just for general relativity). There are three levels of equivalence principles, namely the weak equivalence principle (WEP), the Einstein equivalence principle (EEP), and the strong equivalence principle (SEP). The difference between the SEP and the EEP is that: the EEP (and WEP) only consider the *external* gravitational field of a system (e.g., an elevator) but do not consider the *self*-gravitational field generated by the objects in the system, i.e., they only consider the passive aspects of gravity but ignore the

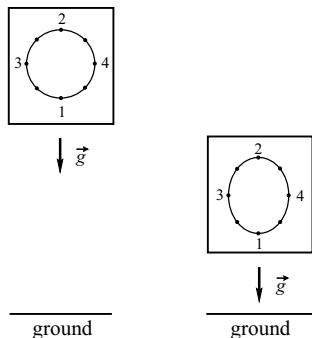
⁸ However, this rule will lead to an ambiguity in the order of the operators when two derivative operators act successively; other considerations need to be taken into account to overcome this issue (See Sect. 7.2). Therefore, the claim that equivalence principles “can carry all the special relativistic laws of physics to curved spacetime” seems to be too strong. Misner et al. (1973) pp. 390–391 has a specific discussion on this.

active aspects; however, the SEP considers both the active and passive aspects, and talks about “self-gravitational systems” which includes from the self-gravity of stars all the way to the gravity between two lead balls in the Cavendish experiment. The EEP can be regarded as the special case of the SEP when self-gravity is negligible. The experimental verification for these three equivalence principles are significant for choosing the gravitational theory. Any gravitational theory will satisfy the WEP (because the WEP has been verified by experiments which are more and more precise, no one would like to create a theory that violates the WEP), but this is not true for the EEP and the SEP. Study shows that [see Will (1995) and Will (2014)], if the EEP is true, then only metric theories can be correct. This indicates that if experiments that are more and more precise can verify the EEP, then there will be less and less room for non-metric theories. Further discussion also shows that [still see Will (1995) and Will (2018)], general relativity satisfies the SEP while none of the other known theories (including the Brans-Dicke theory) does. (Unfortunately, this discussion is not a rigorous proof, and thus till now the conclusion above is technically still a conjecture). Therefore, if experiments that are more and more precise can verify the SEP, then general relativity is very likely to be the only correct gravitational theory. Thus, we can see that the experimental verification for the three equivalence principles has very significant theoretical meaning, and these experiments are now under way with higher and higher precision.

7.6 Tidal Forces and the Geodesic Deviation Equation

Proposition 7.5.1 only shows that the components $\Gamma^{\sigma}_{\mu\nu}$ of the Christoffel symbol in the proper coordinate system of a freely falling non-rotating observer vanish on the world line of this observer. Once off the world line, $\Gamma^{\sigma}_{\mu\nu}$ can be nonvanishing. To see the physical effect of this statement, let us consider the following thought experiment. Put eight balls in an elevator into a circular pattern (the plane of the circle is perpendicular to the ground), as shown in the left part of Fig. 7.8. First we will discuss what happens using Newtonian mechanics. Suppose the line that goes through balls 1 and 2 happens to pass through the Earth’s center, and each ball is at rest relative to the other at the beginning. Since the gravitational field at ball 1 is slightly stronger than that at ball 2, the gravitational acceleration of ball 1 is slightly greater than ball 2, and thus the distance between them will gradually increase. A while later, the whole system will look like what is shown in the right part of Fig. 7.8, which is not round anymore. Imagine ball 1 is an observer, they will find that the distance between them and ball 2 increases with time. However, if the eight balls are arranged in a circle in an inertial spaceship in a region without gravity (and relatively at rest at the beginning), then ball 1 (as an observer) will not find that the distance between them and ball 2 has any change. Thus, even for mechanical experiments, an elevator on the Earth’s surface is not completely equivalent to a spaceship in a region without gravity.

Fig. 7.8 The pattern of balls deforms during freely falling



Although these are thought experiments, phenomena with a similar principle can also be found in daily life. One example is the changing of the tides. Now we will have a simplified analysis of this phenomenon using Newtonian mechanics, in order to highlight the essence of this concept. The leading cause of the tidal phenomenon is the Moon, while the Sun gives a secondary contribution. Ignoring the effect of the Sun can simplify the problem a lot without changing the essence of it. The Earth, as an object, is located in the gravitational field of the Moon. Assume that the Earth's surface is covered by a layer of sea water. Consider two points A and B on the water's surface, such that the line going through them passes through the Earth's center. Suppose at a certain moment A is the closest to the Moon, then B is the furthest from the Moon. The gravitational forces from the Moon on A and B are different, so the two points will move away from each other; thus, the sea's surface near A and B will bulge outwards (left, Fig. 7.9).⁹ As the Earth rotates, A will not face the Moon, and the sea level will drop. After the Earth rotates half a cycle, A is the furthest from the Moon (right, Fig. 7.9), and the sea water will rise again. For someone who is freely falling near the ground, the distance from the Earth's center to their head and feet are different, so there also exists a force that stretches their body (if one only considers the Earth's gravitational field), although this "tidal force" is so small that they will not be able to feel it. If you are freely falling at the surface of a neutron star, the tidal force can be as large as 10^{11} N, and you will be torn apart and dead. Remark: ① A neutron star is a celestial object that is composed mainly of neutrons, whose density can be as high as 10^{14} times that of water! The high density causes an extremely high gradient of the surface gravitational field, see Sect. 9.3. ② According to the usual estimation, the critical pressure or tension that a human body can tolerate (above which the body will be torn apart) is about 10^7 N/m².

The discussion above indicates that any object in the gravitational fields of the Earth and the Moon experiences a tidal force. In fact, the tidal phenomenon is a

⁹ From the viewpoint of an observer on the Earth, the reason for A to bulge is the combination of two forces: (a) the Moon's gravitational force, and (b) the centrifugal force caused by the Earth's circular motion around the barycenter of the Earth and the Moon. The net force of these two forces is called the **tide-generating force** (or **tide-raising force**).

Fig. 7.9 Schematic figures for the tidal phenomena

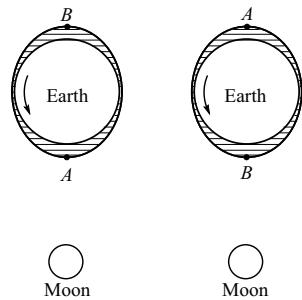
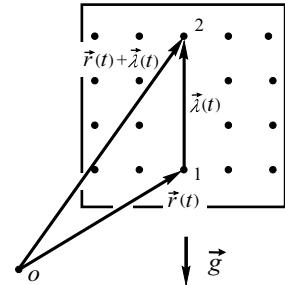


Fig. 7.10 The small balls put everywhere inside an Einstein's elevator



universal feature of gravitational fields. We will discuss the tidal phenomenon quantitatively using Newton's theory of gravity and general relativity, respectively.

First, we use Newton's theory of gravity. Without loss of generality, we still take the example of an Einstein elevator close to the Earth's surface. Suppose there are small balls everywhere inside the elevator (see Fig. 7.10). Let $\vec{r}(t)$ and $\vec{r}(t) + \vec{\lambda}(t)$ represent the position vectors of two balls 1 and 2 next to each other relative to the origin o in a Cartesian coordinate system, then $\vec{\lambda}(t)$ is the position vector of ball 2 relative to ball 1, and hence $d^2\vec{\lambda}/dt^2$ is the acceleration (**tidal acceleration**) of ball 2 relative to ball 1. To calculate the tidal acceleration, one can use the gravitational potential ϕ and Newton's second law to write that

$$\begin{aligned} \frac{d^2x^i}{dt^2} &= -\frac{\partial\phi}{\partial x^i}\Bigg|_{\vec{r}}, \\ \frac{d^2(x^i + \lambda^i)}{dt^2} &= -\frac{\partial\phi}{\partial x^i}\Bigg|_{\vec{r}+\vec{\lambda}} \cong -\frac{\partial\phi}{\partial x^i}\Bigg|_{\vec{r}} - \frac{\partial}{\partial x^j}\frac{\partial\phi}{\partial x^i}\Bigg|_{\vec{r}}\lambda^j, \end{aligned}$$

Subtracting these two equations yields

$$\frac{d^2\lambda^i}{dt^2} = -\frac{\partial^2\phi}{\partial x^i\partial x^j}\Bigg|_{\vec{r}}\lambda^j, \quad i = 1, 2, 3, \quad (7.6.1)$$

which is the expression for the tidal acceleration in Newton's theory of gravity.

Using (7.6.1) one can have a clear idea about the change of the distance between the two balls in Fig. 7.8. Choose a coordinate system $\{x, y, z\}$ such that the z -axis is pointing straight upward, then the z -component of the relative acceleration between balls 1 and 2 is

$$\tilde{a}^z \equiv \frac{d^2 \lambda^z}{dt^2} = -\frac{d^2 \phi}{dz^2} \lambda^z = -\frac{d^2 \phi}{dr^2} \lambda^z, \quad (7.6.2)$$

where r is the distance between ball 1 and the Earth's center. The Earth's gravitational potential is $\phi_{\oplus} = -GM_{\oplus}/r_{\oplus}$, and hence $\tilde{a}^z = 2GM_{\oplus}\lambda^z/r_{\oplus}^3$. Suppose the initial distance between the two balls is $\lambda^z = 1$ m. Plugging in the following numerical values in SI: $G = 6.67 \times 10^{-11}$, $M_{\oplus} = 6 \times 10^{24}$, $r_{\oplus} = 6.37 \times 10^6$ yields $\tilde{a}^z = 0.31 \times 10^{-5}$ m·s⁻². Suppose the two balls are initially at rest with respect to each other, then the increment of their distance after $\Delta t = 5$ s will be

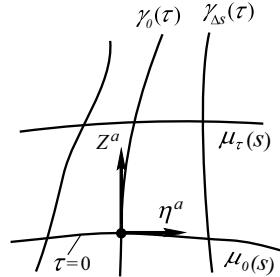
$$\Delta\lambda^z = \frac{1}{2}\tilde{a}^z(\Delta t)^2 = \frac{1}{2} \times 0.31 \times 10^{-5} \times 5^2 \cong 4 \times 10^{-5} \text{ m}. \quad (7.6.3)$$

Now we will investigate the tidal phenomenon from the perspective of general relativity. What we will show is that the tidal phenomenon is an inevitable outcome of the intrinsic curvature of spacetime. Again we will take Fig. 7.10 as an example. Each ball can be viewed as a freely falling observer, whose world line is a timelike geodesic with the proper time τ as the affine parameter. These geodesics form a **geodesic congruence** in an open subset U of the spacetime¹⁰ (physically it corresponds to a freely falling reference frame), and the tangent vectors $Z^a \equiv (\partial/\partial\tau)^a$ of the geodesics form a timelike vector field on U . Let $\mu_0(s)$ be a smooth transverse curve¹¹ [transverse means that the tangent vector of any point on $\mu_0(s)$ is not tangent to the geodesic passing through this point], then each geodesic $\gamma(\tau)$ in the congruence that intersects $\mu_0(s)$ can be labeled by s , i.e., it can be denoted by $\gamma_s(\tau)$, in which s is the value of s at the intersection of this geodesic and $\mu_0(s)$. Choose the initial setting of the proper time of each $\gamma_s(\tau)$ such that the τ at the intersection of $\mu_0(s)$ and each $\gamma_s(\tau)$ is zero. Suppose ϕ_τ is an element of the one-parameter (local) group of diffeomorphisms corresponding to the vector field Z^a and $\mu_\tau(s)$ represents the image of the curve $\mu_0(s)$ under the map $\phi(\tau)$ (see Fig. 7.11). All the curves $\mu_\tau(s)$ with different values of τ cover a subset of \mathcal{S} , on which each point is determined by two real numbers (coordinates) τ and s , and therefore \mathcal{S} is a two dimensional manifold. All the geodesics on \mathcal{S} forms a subset of the geodesic congruence, where each geodesic can be labeled using a parameter s , and thus this subset is also called a one-parameter family of geodesics. (The geodesics in the congruence fill a 4-dimensional open subset U of the spacetime, while this one-parameter family of geodesics only covers a 2-dimensional surface \mathcal{S}). In conclusion, a given transverse curve $\mu_0(s)$ picks a one-parameter family of geodesics $\{\gamma_s(\tau)\}$. Let $\eta \equiv (\partial/\partial s)^a$, then Z^a and η^a are the coordinate basis vector fields of \mathcal{S} , and hence they commute:

¹⁰ A **congruence of curves** in U is a family of curves, such that for each $p \in U$ there is a unique curve in this family passing through p .

¹¹ Some other conditions also need to be satisfied (e.g., non-self-intersecting).

Fig. 7.11 A transverse curve picks a one-parameter family of geodesics $\{\gamma_s(\tau)\}$, the paper represents the 2-dimensional surface spanned by this family



$$0 = [Z, \eta]^a = Z^b \nabla_b \eta^a - \eta^b \nabla_b Z^a, \quad (7.6.4)$$

where ∇_a can be any torsion-free derivative operator. Choose the ∇_a associated with the spacetime metric, then

$$\begin{aligned} Z^b \nabla_b (\eta^a Z_a) &= \eta_a Z^b \nabla_b Z^a + Z_a Z^b \nabla_b \eta^a = Z_a Z^b \nabla_b \eta^a \\ &= Z_a \eta^b \nabla_b Z^a = \frac{1}{2} \eta^b \nabla_b (Z_a Z^a) = 0, \end{aligned} \quad (7.6.5)$$

where the second equality used the fact that Z^a is the tangent vector of the geodesic, the third equality used (7.6.4) and the fifth equality used the fact that $Z_a Z^a = -1$ at each point. Equation (7.6.5) indicates that $\eta^a Z_a$ is a constant along any geodesic $\gamma_s(\tau)$. Therefore, as long as we choose $\mu_0(s)$ in the first place such that it is orthogonal to all $\gamma_s(\tau)$ (which is always possible), then any $\mu_\tau(s)$ will be orthogonal to $\gamma_s(\tau)$. After this choice, the η^a of each point on \mathcal{S} can be viewed as a spatial vector of the geodesic observer $\gamma_s(\tau)$ passing through this point, and thus from now on we will denote η^a by w^a in this text. Suppose Δs is small, then $\gamma_0(\tau)$ and $\gamma_{\Delta s}(\tau)$ can be viewed as the world lines of ball 1 and ball 2 in Fig. 7.10, respectively. Now we call $\gamma_0(\tau)$ the **fiducial observer**, and set $\lambda^a \equiv w^a \Delta s$, then λ^a can be regarded as the λ in Fig. 7.10, namely the position vector of ball 2 relative to the fiducial observer (ball 1). Hence, $\tilde{u}^b \equiv Z^a \nabla_a \lambda^b$ can now be interpreted as the 3-velocity of ball 2 relative to the fiducial observer. [Note that it is a spatial vector field on the world line $\gamma_0(\tau)$ of ball 1 since $Z_b (Z^a \nabla_a \lambda^b) = Z^a \nabla_a (Z_b \lambda^b) - \lambda^b Z^a \nabla_a Z_b = 0$, where in the second equality we used the geodesic equation $Z^a \nabla_a Z_b = 0$ and the fact that λ^b is spatial (i.e., $Z_b \lambda^b = 0$)]. Similarly, $\tilde{a}^c \equiv Z^a \nabla_a (Z^b \nabla_b \lambda^c)$ can be interpreted as the 3-acceleration of ball 2 relative to ball 1 [which is also a spatial vector field on $\gamma_0(\tau)$]. Consider a third geodesic $\gamma_{\Delta \bar{s}}(\tau)$ in the one-parameter family of geodesics (which corresponds to a ball $\bar{2}$ next to ball 2 on the line passing through 1 and 2). The position vector $\bar{\lambda}^a$ of it relative to ball 1 is naturally $\bar{\lambda}^a = w^a \Delta \bar{s}$, and hence the ratio of the tidal accelerations of ball $\bar{2}$ and ball 2 is a constant $\Delta \bar{s}/\Delta s$. Thus, instead of considering specific balls 2, $\bar{2}$, etc. (i.e., using λ^a), we can directly use w^a to define the following universal quantities which apply to all the balls close to ball 1 in the one-parameter family of geodesics:

$$u^b := Z^a \nabla_a w^b, \quad (7.6.6)$$

$$a^c := Z^a \nabla_a u^c = Z^a \nabla_a (Z^b \nabla_b w^c), \quad (7.6.7)$$

both of which are spatial vector fields living on the fiducial geodesic $\gamma_0(\tau)$. In fact, w^a plays the role of a measuring unit of the position vectors of this family: the position vector of any $\gamma_{\Delta s}(\tau)$ is equal to w^a times Δs . Note that w^a has different names in different works, we refer it to as the **separation vector**, which is in agreement with Misner et al. (1973) and Hawking and Ellis (1973). Similarly, u^b and a^c also play the roles of measuring units for the 3-velocity and 3-acceleration of this family, which are called the **3-velocity** and the **3-acceleration (tidal acceleration)** measured by ball 1, respectively. Given a one-parameter family of geodesics and a fiducial geodesic $\gamma_0(\tau)$ in the family, a 3-velocity field u^b and a 3-acceleration field a^c will be determined. Our mission is to reveal the close relationship between a^c and the spacetime curvature, see the following proposition:

Proposition 7.6.1 *The tidal acceleration measured by an arbitrary fiducial geodesic $\gamma_0(\tau)$ in any one-parameter family of timelike geodesics has the following relation with the spacetime curvature tensor [called the **geodesic deviation equation**] :*

$$a^c = -R_{abd}{}^c Z^a w^b Z^d. \quad (7.6.8)$$

Proof

$$a^c = Z^a \nabla_a (Z^b \nabla_b w^c) = Z^a \nabla_a (w^b \nabla_b Z^c) = w^b Z^a \nabla_a \nabla_b Z^c + (Z^a \nabla_a w^b) \nabla_b Z^c = p^c + q^c, \quad (7.6.9)$$

[in the second step we used (7.6.4), i.e., $[Z, w]^b = 0$] where $p^c \equiv w^b Z^a \nabla_a \nabla_b Z^c$, and $q^c \equiv (Z^a \nabla_a w^b) \nabla_b Z^c$. Also,

$$\begin{aligned} p^c &= w^b Z^a \nabla_b \nabla_a Z^c - w^b Z^a R_{abd}{}^c Z^d = w^b \nabla_b (Z^a \nabla_a Z^c) - (w^b \nabla_b Z^a) \nabla_a Z^c - R_{abd}{}^c Z^a w^b Z^d \\ &= -(Z^b \nabla_b w^a) \nabla_a Z^c - R_{abd}{}^c Z^a w^b Z^d = -q^c - R_{abd}{}^c Z^a w^b Z^d, \end{aligned}$$

where in the third equality we used the geodesic equation and (7.6.4). Plugging the above equation into (7.6.9) yields (7.6.8). \square

Now we will make a few more comments on the geodesic deviation equation.

(1) The geodesic deviation equation (7.6.8) is an equation that describes the relative acceleration a^c between two neighboring (“infinitesimally nearby”) geodesics, and a^c is the second order derivative of the separation vector w^a that describes the separation of the two curves. Surely there will be a separation between the two curves ($w^a \neq 0$), and the separation vector may change with time ($u^a \neq 0$), but there is not necessarily a deviation (a^c is not necessarily nonvanishing).¹²

¹² There exists such geodesic families in flat spacetime, in which we have $u^b = 0$ and $a^c = 0$ on a fiducial geodesic $\gamma_0(\tau)$ (such as a parallel geodesic family). There also exists such geodesic families in flat spacetime, where we have $u^b \neq 0$ on $\gamma_0(\tau)$ [one can just let $\gamma_0(\tau)$ and the nearby geodesic become not parallel]. However, there does not exist such a geodesic family, where $a^c \neq 0$ on $\gamma_0(\tau)$ unless the spacetime is not flat.

(2) Equation (7.6.8) reflects the close relationship between a^c and the spacetime curvature tensor $R_{abc}{}^d$: for flat spacetime ($R_{abc}{}^d = 0$), a^c must vanish, and thus the geodesics that are initially parallel will always be parallel [see the footnote after (1)]. However, as long as $R_{abc}{}^d \neq 0$, there will exist a geodesic family whose geodesic deviation (characterized by a^c) is nonvanishing, this is reflected by the fact that the geodesics that are initially parallel will eventually no longer be parallel. The precise meaning of “initially parallel” is that $u^b|_{\tau=0} \equiv Z^a \nabla_a w^b|_{\tau=0} = 0$. This equation indicates that, by means of the physical meaning of u^b with respect to a timelike geodesic family, the relative 3-velocity between two neighboring geodesics is zero at the beginning ($\tau = 0$), and hence is said to be “initially parallel”. However, as long as $a^c|_{\tau=0} \equiv Z^b \nabla_b u^c|_{\tau=0} \neq 0$, after a while u^b will not be zero anymore, i.e., the two geodesics will “become not parallel”. Just as we said in Sect. 3.5, one of the equivalent formulations for the curvature tensor being nonvanishing is that there exist geodesics that are parallel at first which become not parallel.

(3) The Christoffel symbols $\Gamma^\sigma{}_{\mu\nu}$ depend on the coordinate system. By choosing the proper coordinate system of a freely falling non-rotating observer one can make the Christoffel symbols vanish on the world line of the observer (see Proposition 7.5.1), and this can account for the weightlessness of the observer in Einstein’s elevator. However, the tidal acceleration a^c is directly related to the Riemann tensor $R_{abc}{}^d$ (7.6.8), and as a tensor, the latter cannot be made to vanish by choosing any coordinate system. Thus, the tidal acceleration cannot be eliminated by a coordinate transformation. Although the observer in Einstein’s elevator cannot feel gravity (the “gravitational field strength” at this observer is zero), they can still feel the tidal force. This is an interpretation of Fig. 7.8 from general relativity. On the other hand, at least indirectly, the $\Delta\lambda^z$ in (7.6.3) being small verifies the statement that “the effect of spacetime curvature is only manifested in a spacetime region which is large enough.”

(4) So far we only focused on timelike geodesic families. We choose the proper time as the affine parameter, and choose $\mu_\tau(s)$ to be orthogonal to the geodesics. This is for no reason except to emphasize the physical meaning that a^c is the tidal acceleration (in order to have a better correspondence with Fig. 7.10). From the pure mathematical perspective, the geodesic deviation equation (7.6.8) also holds for spacelike and null geodesic families, one just needs to interpret τ as the affine parameter of a geodesic. In this case a^c no longer has the physical interpretation of the tidal acceleration, and the separation vector η^a does not need to be orthogonal to Z^a . Actually, the geodesic deviation equation also holds for a metric with a non-Lorentzian signature. Furthermore, one can even talk about geodesics on a manifold without a metric as long as there is a derivative operator; although orthogonality is not defined, there is still a geodesic deviation equation, i.e., we have the following proposition:

Proposition 7.6.1' *The geodesic deviation equation of an arbitrary one-parameter family of geodesics $\{\gamma_s(\lambda)\}$ in (M, ∇_a) is*

$$a^c = -R_{abd}{}^c T^a \eta^b T^d , \quad (7.6.8')$$

where R_{abd}^c is the Riemann tensor, $T^a \equiv (\partial/\partial\lambda)^a$ is the tangent vector of the fiducial geodesic $\gamma_0(\lambda)$, η^a is the separation vector on $\gamma_0(\lambda)$ (as defined before), and $a^c \equiv T^a \nabla_a (T^b \nabla_b \eta^c)$.

Proof The same as the proof of Proposition 7.6.1. \square

[Optional Reading 7.6.1]

The tidal acceleration a^c in (7.6.8) is defined in terms of w^a (7.6.7). To compare with Newton's theory of gravity, we introduced $\lambda^a \equiv w^a \Delta s$ and considered it as corresponding to the relative position vector $\vec{\lambda}$. Why can λ^a be interpreted as the position vector of ball 2 relative to ball 1? Suppose p and q are two arbitrary points in flat space, w'^a is the unit tangent vector of the line between p and q at p , and $\Delta s'$ is the length of the line between the two points, then $\lambda^a \equiv w'^a \Delta s'$ can be referred to as the position vector of p relative to q (note that $|\lambda^a| = \Delta s'$). Back to the problem of the geodesic deviation in curved space. Take any $\mu_\tau(s)$ in the family of transverse curves, and let $p \equiv \mu_\tau(0)$, $q \equiv \mu_\tau(\Delta s)$ (p and q represent the fiducial observer and the point mass being measured at a time τ , respectively). Use the arc length s' to reparametrize $\mu_\tau(s)$, i.e., $\mu'_\tau(s') = \mu_\tau(s)$, and let w^a and w'^a be the tangent vectors of $\mu_s(s)$ and $\mu'_{\tau'}(s')$ at p , respectively, then $w'^a = w^a ds/ds'$. Hence, if we set $\lambda^a \equiv w^a \Delta s$, then we have $\lambda \equiv w^a \Delta s = w'^a \Delta s'$ when Δs is small. Noticing that $|w'^a| = 1$, we see that $|\lambda^a| = \Delta s'$; comparing with the position vector in flat space, we may say that λ^a is the position vector of ball 2 relative to ball 1. In the main text one does not need to introduce the arc length parameter s' , and thus there is no w'^a , so one just needs to care about w^a whose length changes with τ (note that Δs does not change with τ). The change of the “distance” between balls 1 and 2 is completely manifested by the change of $|w^a|$ with τ , and so defining the relative 3-velocity and relative 3-acceleration using $\lambda^a \equiv w^a \Delta s$ has a perfect correspondence with Fig. 7.10.

[The End of Optional Reading 7.6.1]

[Optional Reading 7.6.2]

If we add a constant related to s to the τ of each $\gamma_s(\tau)$, then $\mu_\tau(s)$ will become non-orthogonal to the geodesics, and thus whether or not η^a and Z^a are orthogonal depends on the zero setting of the proper time of each geodesic. Further, what if we take an arbitrary affine parameter τ' to substitute for τ ? Since τ is an affine parameter, it follows from Theorem 3.3.3 that τ' is an affine parameter if and only if $\tau' = \alpha\tau + \beta$. α and β should of course be constants on each geodesic (and $\alpha \neq 0$), but they can be different for different geodesics, i.e., α and β can be functions of s : $\tau' = \alpha(s)\tau + \beta(s)$. This change of the affine parameter can be viewed as a coordinate transformation $\{\tau, s\} \mapsto \{\tau', s'\}$ on the 2-dimensional manifold \mathcal{S} , where

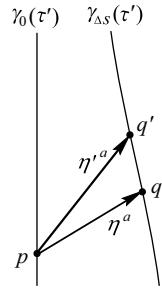
$$s' = s, \quad \tau' = \alpha(s)\tau + \beta(s). \quad (7.6.10)$$

Let Z'^a and η'^a represent the new coordinate basis vectors, i.e., $Z'^a \equiv (\partial/\partial\tau')^a$ and $\eta'^a \equiv (\partial/\partial s')^a$, then it is not difficult to show that

$$Z'^a = \alpha^{-1} Z^a, \quad \eta'^a = \eta^a + v Z'^a, \quad (7.6.11)$$

where $v(\tau, s) \equiv -(\tau d\alpha/ds + d\beta/ds)$ can be viewed as a function on \mathcal{S} . Since we only care about the separation between the fiducial geodesic $\gamma_0(\tau')$ and a geodesic $\gamma_{\Delta s}(\tau')$ next to it, η'^a and η^a can be viewed as vectors describing the same separation (Fig. 7.12). That is, if the separation vectors η'^a and η^a only differ by multiplication by a factor, they describe the same separation. Thus, there exists a “gauge arbitrariness” on the choice of the separation vector. If one insists to use the proper time, but allows each geodesic to have arbitrary zero setting, this is equivalent to setting $\alpha = 1$ in (7.6.10) while letting $\beta(s)$ be arbitrary. Then $Z'^a = Z^a$, $\eta'^a = \eta^a + v Z^a$, and $v = -d\beta/ds$. Equation (7.6.8) can be expressed as

Fig. 7.12 η'^a and η^a describe the same separation



$$a'^c = -R_{abd}{}^c Z'^a \eta'^b Z'^d = -R_{abd}{}^c Z^a (\eta^b + v Z^b) Z^d = a^c,$$

(where we used $R_{abd}{}^c Z^a Z^b Z^d = R_{[ab]d}{}^c Z^{(a} Z^{b)} Z^d = 0$) and thus the zero setting does not affect the value of a^c . However, if one does not insist on using the proper time, i.e., allows $\alpha \neq 1$, then one can only have

$$a'^c = -R_{abd}{}^c Z'^a \eta'^b Z'^d = \alpha^{-2} a^c.$$

This is natural since substituting τ' for the proper time τ is equivalent to substituting a “coordinate clock” for the standard clock. The rate of this coordinate clock is α^{-1} times the rate of the standard clock, and the “tidal acceleration” measured using this clock is naturally α^{-2} times the result measured by the standard clock.

[The End of Optional Reading 7.6.2]

[Optional Reading 7.6.3]

A solution η^b to the geodesic deviation equation (7.6.8') is called a **Jacobi field** on the geodesic $\gamma(\lambda)$ being considered. Two points $p, q \in \gamma(\lambda)$ are said to be **conjugate** if there exists a non-vanishing Jacob field η^b on $\gamma(\lambda)$, which vanishes at p and q . In this case, we also say that p and q are a pair of **conjugate points** on the geodesic $\gamma(\lambda)$. For instance, the south and north poles s and n on the 2-dimensional sphere shown in Fig. 7.13 are a pair of conjugate points on the geodesic γ from s to n (half of the great circle). It is not difficult to accept the following intuitive statement: $p, q \in \gamma$ are a pair of conjugate points if there exists a geodesic from p to q that is infinitesimally close to but different from γ (such as the γ' in the figure). The precise meaning of the condition after the word “if” is: there exists a one-parameter family of geodesics from p to q which includes γ . The logic above can be formulated as:

There exists a geodesic from p to q that is infinitesimally close to but different from γ
 \Leftrightarrow there exists a one-parameter family of geodesics from p to q which includes γ .
 $\Rightarrow p, q \in \gamma$ are a pair of conjugate points \Leftrightarrow there exists a non-vanishing Jacob field η^b on $\gamma(\lambda)$, which vanishes at p and q .

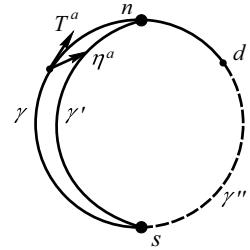
This logic can help us clarify two subtle problems, which will be introduced as follows in the manner of Q&A (we stipulate that γ is a geodesic):

Q: Suppose $p, q \in \gamma$ are a pair of conjugate points, does there exist a geodesic from p to q that is different from but infinitesimally close to γ ?

A: Not necessarily. Because the \Rightarrow in the relation above cannot be changed to \Leftrightarrow . There does exist such situations, in which $p, q \in \gamma$ are conjugate but one cannot find a geodesic from p to q that is different from but infinitesimally close to γ (omitted).

Q: Suppose there exists a geodesic γ'' that passes through $p, q \in \gamma$ and is different from γ , can we say that p and q are conjugate?

Fig. 7.13 s and n are a pair of conjugate points, while s and d are not



A: No. Because only γ'' existing cannot guarantee that there exists a one-parameter family of geodesics between p and q which includes γ . A counter example: extending the great arc γ to d , and denote this major arc as $\tilde{\gamma}$, then $s, d \in \tilde{\gamma}$ and there exists a geodesic γ'' (the minor arc of the great circle) which passes s and d and is different from $\tilde{\gamma}$. However, $s, d \in \gamma$ are not a pair of conjugate points since (intuitively speaking) there does not exist a geodesic connecting s and d that is “infinitesimally close” to $\tilde{\gamma}$ (γ'' is certainly not close to it), or (precisely speaking) there does not exist a nonvanishing Jacobi field η^b satisfying the vanishing at the end points condition.

For the significance of conjugate points on the arc length problem, see Sect. 3.3; for the use of it in the proofs of the singularity theorems, see Wald (1984) pp. 223–233.

[The End of Optional Reading 7.6.3]

7.7 The Einstein Field Equation

Since the distribution of matter produces gravity, and gravity is manifested by the spacetime curvature, a natural hypothesis is that the spacetime curvature is affected by the matter distribution. The matter distribution is described by the energy-momentum tensor T_{ab} , and hence there should exist an equation that relates T_{ab} and the spacetime curvature. Considering that Newton’s theory of gravity should be the weak-field and low-speed approximation of general relativity, the comparison between the geodesic deviation equation (7.6.8) and the tidal force acceleration (7.6.1) in Newton’s theory of gravity provides important clues for seeking (guessing) this equation. Since the a^c in (7.6.8) is defined in terms of w^a instead of λ^a , for convenience’s sake, we should change the λ^i in (7.6.1) to w^i . Suppose $\{x^i\}$ is a Cartesian system of the 3-dimensional Euclidean space, then (7.6.1) can be written as

$$\begin{aligned} a^c &= a^i \left(\frac{\partial}{\partial x^i} \right)^c = \left(\frac{\partial}{\partial x^i} \right)^c \frac{d^2 w^i}{dt^2} = - \left(\frac{\partial}{\partial x^i} \right)^c w^j \frac{\partial}{\partial x^j} \left(\frac{\partial \phi}{\partial x^i} \right) \\ &= - \left(\frac{\partial}{\partial x^i} \right)^c w^b \partial_b \left(\frac{\partial \phi}{\partial x^i} \right) = -w^b \partial_b \left[\left(\frac{\partial}{\partial x^i} \right)^c \left(\frac{\partial \phi}{\partial x^i} \right) \right] = -w^b \partial_b \partial^c \phi. \end{aligned}$$

This is the tidal acceleration derived from Newton’s theory of gravity, which should be an approximation of the a^c derived from general relativity. Therefore, the comparison

between the above equation and (7.6.8) implies the following correspondence:

$$R_{abd}{}^c Z^a Z^d \leftrightarrow \partial_b \partial^c \phi . \quad (7.7.1)$$

Contracting the indices b and c yields

$$R_{abd}{}^b Z^a Z^d \leftrightarrow \partial_b \partial^b \phi = \nabla^2 \phi = 4\pi \rho = 4\pi T_{ad} Z^a Z^d ,$$

where $\nabla^2 \phi = 4\pi \rho$ is Poisson's equation in Newton's theory of gravity, and in the last step we used the property 3(a) of T_{ab} in Sect. 6.4 (μ is changed to ρ). From the above correspondence, we expect the following equation to hold:

$$R_{ad} Z^a Z^d = 4\pi T_{ad} Z^a Z^d . \quad (7.7.2)$$

The simplest assumption that satisfies the above equation is

$$R_{ab} = 4\pi T_{ab} . \quad (7.7.3)$$

In fact, this is what Einstein assumed and published initially. However, from Sect. 6.4 we can see that the energy-momentum tensor T_{ab} satisfies $\partial^a T_{ab} = 0$; using the minimal substitution rule in Sect. 7.2 we have $\nabla^a T_{ab} = 0$, and hence (7.7.3) leads to

$$\nabla^a R_{ab} = 0 , \quad (7.7.3')$$

which will lead to physically unacceptable consequences. Contracting the Bianchi identity $\nabla_{[a} R_{bc]d}{}^e = 0$ yields $\nabla_{[a} R_{bc]d}{}^a = 0$, and thus

$$0 = \nabla_a R_{bcd}{}^a + \nabla_c R_{abd}{}^a + \nabla_b R_{cad}{}^a = \nabla_a R_{bcd}{}^a - \nabla_c R_{bd} + \nabla_b R_{cd} .$$

Raising the index d using the metric and contracting it with the lower index b yields

$$0 = \nabla_a R_c{}^a - \nabla_c R + \nabla_b R_c{}^b = 2\nabla^a R_{ca} - \nabla_c R ,$$

and so (7.7.3') requires that

$$\nabla_c R = 0 . \quad (7.7.4)$$

This is an additional condition enforced on R_{ab} by (7.7.3'). To illustrate that this condition is unacceptable, let $T \equiv g^{ab} T_{ab}$, and raise the index b in (7.7.3) and contracting it with the lower index a . Then we have $R = 4\pi T$, and hence (7.7.4) leads to $\nabla_c T = 0$, i.e., T is a constant in the whole matter field. Take a perfect fluid as an example, it follows from (6.5.1) (changing μ to ρ) that

$$T = T_a{}^a = \rho U_a U^a + p(\delta_a{}^a + U_a U^a) = -\rho + 3p .$$

Under the Newtonian approximation we have $\rho \gg p$, and hence $T \cong -\rho$. Therefore, T being a constant means that the proper energy density ρ is a constant in the whole fluid field. This obviously does not agree with the perfect fluid case in physics, and thus (7.7.3) must be modified. The problem is that $\nabla^a T_{ab} = 0$ while $\nabla^a R_{ab}$ should not vanish. If one can find a symmetric tensor G_{ab} of type $(0, 2)$ that only depends on the spacetime geometry, which not only satisfies $\nabla^a G_{ab} = 0$, but can also replace R_{ab} in an equation similar to (7.7.3) which still leads to (7.7.2), then the issue will be solved. This G_{ab} is not difficult to find, since it can be easily seen from the equation above (7.7.4) that

$$0 = \nabla^a R_{ab} - \frac{1}{2} \nabla_b R = \nabla^a R_{ab} - \frac{1}{2} g_{ab} \nabla^a R = \nabla^a (R_{ab} - \frac{1}{2} R g_{ab}),$$

what is inside the parenthesis on the right-hand side of this equation can be taken as G_{ab} . Therefore, we can define G_{ab} as

$$G_{ab} \equiv R_{ab} - \frac{1}{2} R g_{ab}, \quad \nabla^a G_{ab} = 0, \quad (7.7.5)$$

(G_{ab} is called the Einstein tensor, see Definition 3 and Theorem 3.4.8 in Sect. 3.4), and substitute (7.7.3) by the equation $G_{ab} = 8\pi T_{ab}$, namely we assume

$$R_{ab} - \frac{1}{2} R g_{ab} = 8\pi T_{ab}. \quad (7.7.6)$$

The left-hand side of this equation satisfies $\nabla^a (R_{ab} - \frac{1}{2} R g_{ab}) = 0$ automatically, and thus it is compatible with $\nabla^a T_{ab} = 0$. On the other hand, it is not difficult to see that the above equation will return to (7.7.2) under the Newtonian approximation ($T \cong -\rho$) as we want. First, it follows from (7.7.6) that $8\pi T_a^a = R_a^a - \frac{1}{2} \delta_a^a R = R - 2R = -R$, i.e.,

$$R = -8\pi T, \quad (7.7.7)$$

and hence (7.7.6) leads to

$$R_{ab} = 8\pi T_{ab} + \frac{1}{2} g_{ab} R = 8\pi T_{ab} + \frac{1}{2} g_{ab} (-8\pi T) = 8\pi (T_{ab} - \frac{1}{2} g_{ab} T). \quad (7.7.6')$$

Thus,

$$\begin{aligned} R_{ab} Z^a Z^b &= 8\pi (T_{ab} Z^a Z^b - \frac{1}{2} g_{ab} Z^a Z^b T) = 8\pi (\rho + \frac{1}{2} T) \\ &\cong 8\pi (\rho - \frac{1}{2} \rho) = 4\pi \rho = 4\pi T_{ab} Z^a Z^b, \end{aligned}$$

which is exactly (7.7.2). Therefore, one should take (7.7.6) as the equation describing the relation between the spacetime curvature and a matter field. This equation is dubbed the **Einstein field equation**, which is a basic postulate of general relativity.¹³

In Minkowski space $R_{abc}^d = 0$ everywhere; hence, $G_{ab} = 0$, and from Einstein's equation we know that $T_{ab} = 0$. However, is there any physics if there is no matter? In fact, special relativity studies the motion of physical objects and their interactions, but the gravitational interaction between them is ignored, i.e., the gravitational fields produced by the physical objects are ignored, and therefore the spacetime is approximately flat. Thus, special relativity is the approximation of general relativity when gravity (spacetime curvature) can be ignored. As long as gravity is not negligible, the spacetime cannot be treated as flat, and in principle special relativity cannot be applied.

An important special case is $T_{ab} = 0$, in which Einstein's equation becomes

$$R_{ab} - \frac{1}{2} R g_{ab} = 0, \quad (7.7.8)$$

called the **vacuum Einstein equation**. Given a coordinate system, the components $R_{\mu\nu}$ of the Ricci tensor can be expressed by the components $g_{\mu\nu}$ of the metric and its partial derivatives (up to the second order) [see (3.4.21)], and the dependence of $R_{\mu\nu}$ on $g_{\mu\nu}$ is highly nonlinear.¹⁴ Therefore, (7.7.8) can be viewed as a set of nonlinear 2nd-order partial differential equations for the unknown functions $g_{\mu\nu}$, each solution g_{ab} is a vacuum metric. The Minkowski metric is naturally a solution to the equation (7.7.8), while a solution to (7.7.8) can be a curved metric. An important example is the vacuum solution found by Karl Schwarzschild within two months after the publication of Einstein's equation, see Sect. 8.3 and Chap. 9 for details.

It is not difficult to show that the scalar curvature R vanishes when $T_{ab} = 0$, and thus the vacuum Einstein equation (7.7.8) can be simplified as

$$R_{ab} = 0. \quad (7.7.8')$$

This indicates that the Riemann tensor of a vacuum metric (i.e., a solution to the vacuum Einstein equation) g_{ab} is equal to its Weyl tensor (see Definition 2 of Sect. 3.4), which is usually nonvanishing.

Equation (7.7.6) with $T_{ab} \neq 0$ is called **Einstein's equation with source**, which is similar to Maxwell's equations with source in Minkowski spacetime [see (6.6.10)], except there is an important difference. For Maxwell's equations, one can solve for the unknown F_{ab} when the source (4-current density J^a) is assigned. It seems that for Einstein's equation one can also assign T_{ab} (as a given quantity) and then solve

¹³ The story being told here is a cleaned up version of the much more convoluted path which Einstein actually followed originally. In fact, Einstein did not define the Einstein tensor first, and the form of his equation published in November 1915 was (7.7.6') instead of (7.7.6).

¹⁴ Specifically, the dependence of $G_{\mu\nu}$ on the second order derivatives of $g_{\mu\nu}$ is linear, while the dependence on the first order derivatives is quadratic. What is worse, $G_{\mu\nu}$ also contains the inverse $g^{\mu\nu}$ of $g_{\mu\nu}$ (for raising the indices), which is very complex when expressed as a function of $g_{\mu\nu}$.

for the unknown quantity g_{ab} ; however, there is an issue: T_{ab} is not meaningful when g_{ab} is undetermined. Take a perfect fluid with zero pressure (dust) as an example. To define a dust as a matter field, we mean to assign a 4-velocity field U^a and a proper density field ρ to it. The energy-momentum tensor of the dust is $T_{ab} = \rho U_a U_b$, where $U_a \equiv g_{ac} U^c$. Therefore, as long as g_{ac} is undetermined, the value of T_{ab} is not known. Moreover, the 4-velocity field U^a should be timelike and normalized, and both of these concepts involve the metric g_{ab} , and so one can hardly view U^a as a given quantity when g_{ab} is unknown. Thus, it is improper to treat g_{ab} and T_{ab} as respectively unknown and given quantities. The source of this difference between Einstein's equation and Maxwell's equations is that: the spacetime background (Minkowski spacetime) is already stipulated in Maxwell's theory, and the right-hand side of the equation $\partial^a F_{ab} = -4\pi J_b$ will be a given quantity $-4\pi \eta_{bc} J^c$ when a 4-current vector J^a is given; for Einstein's equation, however, g_{ab} that describes the spacetime background is yet to be determined, and unfortunately, it appears on both sides of the equation, and thus one cannot simply consider the right-hand side as being given beforehand. When solving Einstein's equation, one should treat g_{ab} and the quantities describing matter fields (e.g., for a dust they are U^a and ρ) together as unknown quantities and solve for them simultaneously. We will provide an example of solving Einstein's equation in Sect. 8.4, where the “matter field” will be an electromagnetic field.¹⁵

The non-linearity of Einstein's equation means that it does not satisfy the superposition principle, which leads to many consequences. For instance, the sum of two solutions to an equation is not a solution. This is another significant difference between Einstein's equation and Maxwell's equations.

The Einstein tensor satisfies $\nabla^a G_{ab} = 0$ [see (7.7.5)], and therefore Einstein's equation contains $\nabla^a T_{ab} = 0$, which includes a lot of information about the motion of matter. In fact, for a perfect fluid, this is the equation of motion for the matter field (see Sect. 6.5). For a perfect fluid with zero pressure, i.e., a dust, it follows from $\nabla^a T_{ab} = 0$ that the world line of a dust particle is a geodesic [see (6.5.8) and a few sentences after that]. This conclusion can also be generalized to any object whose self-gravity is weak enough [Fock (1939); Geroch and Jang (1975)]. Thus, the postulate in Sect. 7.1 about the world lines of free particles being geodesics is no longer an independent postulate.

Another completely different approach to obtain Einstein's field equation is through the Lagrangian formulation of general relativity, which will be introduced in Chap. 16 (Volume III). Since it does not involve any knowledge that has not been covered so far, readers who want to learn about deriving Einstein's equation through the variational principle may refer to Sect. 16.1 (except for the optional reading) directly after reading this section.

¹⁵ Conventionally, an electromagnetic field is not classified as a matter field, but as the source of a gravitational field we will later on refer to it as a matter field for convenience.

7.8 Linear Approximation and the Newtonian Limit

7.8.1 Linearized Theory of Gravity

The non-linearity of the Einstein field equation brings many difficulties to the task of solving the equation as well as the study of general relativity in general. In most of the cases the gravitational field is weak, and one can approximate the field equation as a linear equation, which will significantly simplify the problem. In the 4-dimensional language, a weak gravitational field means that the spacetime metric g_{ab} is close to the Minkowski metric η_{ab} .¹⁶ Define γ_{ab} using the following equation:

$$g_{ab} = \eta_{ab} + \gamma_{ab}, \quad (7.8.1)$$

then γ_{ab} is “small”, which means that the components of γ_{ab} in a Lorentzian coordinate system of η_{ab} satisfy $|\gamma_{\mu\nu}| \ll 1$, so that the second and higher order terms can all be neglected. Under this approximation, γ_{ab} can be treated as some kind of physical field (similar to the electromagnetic field) in Minkowski spacetime. The difference between γ_{ab} and an ordinary physical field is that the sum of γ_{ab} and η_{ab} gives the spacetime metric. From this perspective (plus the fact that γ_{ab} is “small”), γ_{ab} can be viewed as a perturbation of η_{ab} . For convenience and to avoid confusion, we stipulate that the tensor indices are all raised and lowered by η^{ab} and η_{ab} (instead of g^{ab} and g_{ab}), with only one exception, which is g^{ab} . g^{ab} will still represent the inverse of g_{ab} rather than $\eta^{ac}\eta^{bd}g_{cd}$. Under the linear approximation, it is not difficult to see from (7.8.1) that

$$g^{ab} = \eta^{ab} - \gamma^{ab}, \quad (7.8.2)$$

as from this we have $g^{ab}g_{bc} = \delta^a_c$ – (second-order terms in γ). Suppose ∂_a and ∇_a are the derivative operators associated with η_{ab} and g_{ab} , respectively, then from (3.2.10) we know that the Christoffel symbol (i.e., the “difference” between ∂_a and ∇_a) in a Lorentzian system is

$$\Gamma^c{}_{ab} = \frac{1}{2}g^{cd}(\partial_a g_{bd} + \partial_b g_{ad} - \partial_d g_{ab}). \quad (7.8.3)$$

Plugging (7.8.1) and (7.8.2) into the above equation and only keeping the first-order terms in γ_{ab} , we have

$$\Gamma^{(1)c}{}_{ab} = \frac{1}{2}\eta^{cd}(\partial_a \gamma_{bd} + \partial_b \gamma_{ad} - \partial_d \gamma_{ab}). \quad (7.8.4)$$

¹⁶ In the linearized theory of gravity, people usually discuss the spacetime with the background manifold \mathbb{R}^4 , or a spacetime region where a flat Lorentzian metric $\tilde{\eta}_{ab}$ can be defined. In the former case the Minkowski metric η_{ab} is globally defined, and in the latter case it is convenient to denote the (locally) flat metric $\tilde{\eta}_{ab}$ as η_{ab} .

Using the property that $\Gamma^{(1)c}_{ab}$ itself is a first-order small term, plugging the above equation into (3.4.20) yields the first-order approximation of the Riemann tensor (with lower indices) of g_{ab} (called the **linearized Riemann tensor**)

$$R_{acbd}^{(1)} = \partial_d \partial_{[a} \gamma_{c]} b - \partial_b \partial_{[a} \gamma_{c]} d . \quad (7.8.5)$$

Using η^{cd} to raise and contract the indices, we obtain the first-order approximation of the Ricci tensor of g_{ab} (the linearized Ricci tensor)

$$R_{ab}^{(1)} = \partial^c \partial_{(a} \gamma_{b)c} - \frac{1}{2} \partial^c \partial_c \gamma_{ab} - \frac{1}{2} \partial_a \partial_b \gamma , \quad (7.8.6)$$

where $\gamma \equiv \gamma^a{}_a = \eta^{ab} \gamma_{ab}$. From this one can easily get the first-order approximation of the Einstein tensor (called the **linearized Einstein tensor**)

$$G_{ab}^{(1)} = R_{ab}^{(1)} - \frac{1}{2} \eta_{ab} R^{(1)} = \partial^c \partial_{(b} \gamma_{a)c} - \frac{1}{2} \partial^c \partial_c \gamma_{ab} - \frac{1}{2} \partial_a \partial_b \gamma - \frac{1}{2} \eta_{ab} (\partial^c \partial^d \gamma_{cd} - \partial^c \partial_c \gamma) . \quad (7.8.7)$$

Therefore,

$$\partial^c \partial_{(a} \gamma_{b)c} - \frac{1}{2} \partial^c \partial_c \gamma_{ab} - \frac{1}{2} \partial_a \partial_b \gamma - \frac{1}{2} \eta_{ab} (\partial^c \partial^d \gamma_{cd} - \partial^c \partial_c \gamma) = 8\pi T_{ab} \quad (7.8.8)$$

is called the **linearized Einstein equation**. Let

$$\bar{\gamma}_{ab} \equiv \gamma_{ab} - \frac{1}{2} \eta_{ab} \gamma , \quad (7.8.9)$$

then the linearized Einstein equation can be further simplified as

$$-\frac{1}{2} \partial^c \partial_c \bar{\gamma}_{ab} + \partial^c \partial_{(a} \bar{\gamma}_{b)c} - \frac{1}{2} \eta_{ab} \partial^c \partial^d \bar{\gamma}_{cd} = 8\pi T_{ab} . \quad (7.8.8')$$

The left-hand side of this equation vanishes when $\partial^b \equiv \eta^{bc} \partial_c$ acts on it, and thus the equation above assures $\partial^b T_{ab} = 0$. This has an important physical meaning: it indicates that the divergence of the energy-momentum tensor vanishes in the linearized theory of gravity, and hence assures that the laws of conservation of energy, momentum and angular momentum also hold in the linearized theory of gravity (as a physical theory).

Equation (7.8.8') can also be further simplified. In order to do this, we first review a heuristic example. Maxwell's equation $\partial^a F_{ab} = -4\pi J_b$ in Minkowski spacetime can be expressed using the electromagnetic 4-potential A_a as [see (6.6.30)]

$$\partial^a \partial_a A_b - \partial_b \partial^a A_a = -4\pi J_b . \quad (7.8.10)$$

Suppose χ is an arbitrary scalar field, then the following transformation for A_a :

$$\tilde{A}_a = A_a + \partial_a \chi \quad (7.8.11)$$

is called a gauge transformation since \tilde{A}_a and A_a correspond to the same F_{ab} . One can always choose χ so that the 4-potential satisfies the Lorenz gauge:

$$\partial^a A_a = 0, \quad (7.8.12)$$

then (7.8.10) can be simplified as

$$\partial^a \partial_a A_b = -4\pi J_b. \quad (7.8.13)$$

In the linearized theory of gravity, there exists a very similar gauge freedom. Suppose ξ^a is an infinitesimal vector field (“infinitesimal” means that the components ξ^μ of ξ^a are small enough so that the product with γ_{ab} or itself can be regarded as second-order terms and neglected), the following transformation of γ_{ab} :

$$\tilde{\gamma}_{ab} = \gamma_{ab} + \partial_a \xi_b + \partial_b \xi_a \quad (7.8.14)$$

is called a **gauge transformation in the linearized theory of gravity**, since it is not difficult to verify from the commutativity of ∂_a and ∂_b that $\eta_{ab} + \tilde{\gamma}_{ab}$ and $\eta_{ab} + \gamma_{ab}$ have the same linearized Riemann tensor. $R_{abcd}^{(1)}$ being invariant leads to the fact that $R_{ab}^{(1)}$ and $G_{ab}^{(1)}$ are invariant. Therefore, if γ_{ab} is a solution to the linearized Einstein equation, then $\tilde{\gamma}_{ab}$ will also be one. This gauge invariance allows us to choose an appropriate γ_{ab} among all the equivalent ones (i.e., to choose an appropriate gauge) to simplify the linearized Einstein equation (7.8.8). As an analogue of the electromagnetic Lorenz gauge condition (7.8.12), we will show below that there exists a subclass in the equivalence class, in which the $\bar{\gamma}_{ab}$ of each γ_{ab} satisfies the following equation:

$$\partial^b \bar{\gamma}_{ab} = 0, \quad (7.8.15)$$

called the **Lorenz gauge condition** of the linearized theory of gravity.¹⁷ From the equation above we can see that the second and third terms on the right-hand side of the linearized Einstein equation (7.8.8') of this type of $\bar{\gamma}_{ab}$ vanish, and hence the equation can be simplified as

$$\partial^c \partial_c \bar{\gamma}_{ab} = -16\pi T_{ab}, \quad (7.8.16)$$

which is very similar to (7.8.13)! Now we will show that (7.8.15) can always be satisfied by choosing ξ^a . Suppose that $\bar{\gamma}_{ab}$ does not satisfy (7.8.15), in order to choose ξ^a such that $\tilde{\gamma}_{ab}$ determined by (7.8.14) has a corresponding

¹⁷ Also called the de Donder gauge condition or harmonic gauge condition of the linearized theory of gravity.

$$\bar{\tilde{\gamma}}_{ab} = \tilde{\gamma}_{ab} - \frac{1}{2}\eta_{ab}\tilde{\gamma} \quad (\tilde{\gamma} \equiv \eta^{ab}\tilde{\gamma}_{ab})$$

that satisfies (7.8.15). A simple calculation starting from (7.8.14) shows that $\partial^b\bar{\tilde{\gamma}}_{ab} = \partial^b\tilde{\gamma}_{ab} + \partial^b\partial_b\xi_a$, and hence as long as we choose a ξ^a satisfying

$$\partial^b\partial_b\xi_a = -\partial^b\tilde{\gamma}_{ab}, \quad (7.8.17)$$

then $\partial^b\bar{\tilde{\gamma}}_{ab} = 0$ is guaranteed. A ξ^a that satisfies (7.8.17) must exist, since the component form of this equation in an inertial coordinate system will be the following familiar equation:

$$-\frac{\partial^2\xi_\mu}{\partial t^2} + \frac{\partial^2\xi_\mu}{\partial x^2} + \frac{\partial^2\xi_\mu}{\partial y^2} + \frac{\partial^2\xi_\mu}{\partial z^2} = -\partial^\nu\bar{\tilde{\gamma}}_{\mu\nu}.$$

When $\bar{\tilde{\gamma}}_{\mu\nu}$ is given, the solutions to it not only exist, but also they are numerous.

[Optional Reading 7.8.1]

There is a subtlety in the derivation from (7.8.3) to (7.8.4) that we should specify. Take the term $g^{cd}\partial_a g_{bd}$ as an example, it can be expressed as

$$g^{cd}\partial_a g_{bd} = (\eta^{cd} - \gamma^{cd})\partial_a\gamma_{bd}, \quad (7.8.18)$$

but why do we only keep $\eta^{cd}\partial_a\gamma_{bd}$? Seeing that the off-diagonal components of η^{cd} vanish but the off-diagonal components of γ^{cd} can be nonzero, why can it still be neglected? This can be interpreted from the perspective of perturbation theory. Consider a one-parameter family of metrics, $g_{ab}(s)$, and a one-parameter family of energy-momentum tensors $T_{ab}(s)$ (with s as the parameter) satisfying

- (a) $G_{ab}(s) = 8\pi T_{ab}(s)$ [where $G_{ab}(s)$ is the Einstein tensor of $g_{ab}(s)$];
- (b) $g_{ab}(0) = \eta_{ab}$, $T_{ab}(0) = 0$;
- (c) There exists a small quantity $\varepsilon > 0$ such that $(g_{ab}(\varepsilon), T_{ab}(\varepsilon))$ is the (g_{ab}, T_{ab}) of the spacetime we are concerned with.

Moreover, we also require that $g_{ab}(s)$ and $T_{ab}(s)$ can both be Taylor expanded:

$$\begin{aligned} g_{ab}(s) &= \eta_{ab} + sg_{ab}^{(1)} + s^2g_{ab}^{(2)} + O(s^3), \\ T_{ab}(s) &= sT_{ab}^{(1)} + s^2T_{ab}^{(2)} + O(s^3). \end{aligned}$$

Plugging the two equations above into $G_{ab}(s) = 8\pi T_{ab}(s)$, and ignoring all the $O(s^2)$ and higher order terms, what we obtain will be the linear (first-order) approximation of the Einstein equation, namely (7.8.8). And the derivation from (7.8.3) to (7.8.4) is one of the steps in this procedure. Since neither γ^{cd} nor $\partial_a\gamma_{bd}$ in (7.8.18) contains a zeroth-order term of s , $\gamma^{cd}\partial_a\gamma_{bd}$ is at least a second-order term. Thus, this term can be neglected and we have (7.8.4).

[The End of Optional Reading 7.8.1]

[Optional Reading 7.8.2]

In the main text above we have introduced gauge transformations in the linearized theory of gravity using the active language. In the passive language, such a transformation is the result of an **infinitesimal coordinate transformation** as follows:

$$x'^\mu = x^\mu - \xi^\mu(x), \quad (7.8.19)$$

(the x in the parentheses is an abbreviation for x^σ) where $\xi^\mu(x)$ are four arbitrary infinitesimal functions of the same order as γ_{ab} . [See Misner et al. (1973) pp. 439–440]. Consider the coordinate components $g_{\rho\sigma} = \eta_{\rho\sigma} + \gamma_{\rho\sigma}$, under the above coordinate transformation the tensor transformation law

$$g'_{\mu\nu}(x') = \frac{\partial x^\rho}{\partial x'^\mu} \frac{\partial x^\sigma}{\partial x'^\nu} g_{\rho\sigma}(x) \quad (7.8.20)$$

can be reduced to

$$\begin{aligned} g'_{\mu\nu}(x') &= \left(\delta^\rho_\mu + \frac{\partial \xi^\rho}{\partial x^\mu} \right) \left(\delta^\sigma_\nu + \frac{\partial \xi^\sigma}{\partial x^\nu} \right) g_{\rho\sigma}(x) \\ &= g_{\mu\nu}(x) + \frac{\partial \xi^\sigma}{\partial x^\nu} g_{\mu\sigma}(x) + \frac{\partial \xi^\rho}{\partial x^\mu} g_{\rho\nu}(x) \\ &= \eta_{\mu\nu} + \gamma_{\mu\nu} + \frac{\partial \xi_\nu}{\partial x^\mu} + \frac{\partial \xi_\mu}{\partial x^\nu}, \quad (\text{where } \xi_\mu = \eta_{\mu\rho} \xi^\rho), \end{aligned}$$

up to terms of higher order than γ_{ab} and ξ^a . Define $\gamma'_{\mu\nu} = g'_{\mu\nu} - \eta_{\mu\nu}$. Then, up to higher order terms,

$$\gamma'_{\mu\nu} = \gamma_{\mu\nu} + \xi_{\mu,\nu} + \xi_{\nu,\mu}.$$

On the other hand, $\gamma'_{\mu\nu}(x) = g'_{\mu\nu}(x') - \eta_{\mu\nu} = [g'_{\mu\nu}(x') - g_{\mu\nu}(x)] + [g_{\mu\nu}(x) - \eta_{\mu\nu}]$ turns out to be

$$\gamma'_{\mu\nu}(x) = [g'_{\mu\nu}(x') - g_{\mu\nu}(x)] + \gamma_{\mu\nu}. \quad (7.8.21)$$

Hence, up to terms of higher order than γ_{ab} and ξ^a , we have $g'_{\mu\nu}(x') - g_{\mu\nu}(x) = \xi_{\mu,\nu} + \xi_{\nu,\mu}$.

To see how the above coordinate description is related to the gauge transformation (7.8.14) in the active language, we consider the one-parameter local group of diffeomorphisms generated by a vector field X^a , denoted by ϕ_λ (see Optional Reading 2.2.2), with λ as the parameter. Here we choose X^a such that the infinitesimal vector field ξ^a in the gauge transformation is $\xi^a = \varepsilon X^a$, where ε is an infinitesimal number. For both g_{ab} and $\tilde{g}_{ab}(\lambda) \equiv \phi_\lambda^* g_{ab}$ we can split them as

$$g_{ab} = \eta_{ab} + \gamma_{ab}, \quad \tilde{g}_{ab}(\lambda) = \eta_{ab} + \tilde{\gamma}_{ab}(\lambda),$$

and obtain that $\eta_{ab} + \tilde{\gamma}_{ab}(\lambda) = \tilde{g}_{ab}(\lambda) = \phi_\lambda^* g_{ab} = \phi_\lambda^* \eta_{ab} + \phi_\lambda^* \gamma_{ab}$, i.e.,

$$\tilde{\gamma}_{ab}(\lambda) = \phi_\lambda^* \gamma_{ab} + \phi_\lambda^* \eta_{ab} - \eta_{ab}.$$

When λ is small, one can rewrite the above equation by means of Lie derivatives as

$$\tilde{\gamma}_{ab}(\lambda) = \gamma_{ab} + \lambda \mathcal{L}_X \gamma_{ab} + \lambda \mathcal{L}_X \eta_{ab} + O(\lambda^2) = \gamma_{ab} + \mathcal{L}_{\lambda X} \gamma_{ab} + \mathcal{L}_{\lambda X} \eta_{ab} + O(\lambda^2),$$

where the last step can be easily seen from (4.2.8). Ignoring the higher order terms, we have

$$\tilde{\gamma}_{ab} \equiv \tilde{\gamma}_{ab}(\varepsilon) = \gamma_{ab} + \mathcal{L}_\xi \eta_{ab} = \gamma_{ab} + \partial_a \xi_b + \partial_b \xi_a,$$

where we have set $\lambda = \varepsilon$, and (4.3.1') is used in the last step. Therefore, the gauge transformation (7.8.14) can be obtained from changing the metric g_{ab} to $\tilde{g}_{ab}(\varepsilon)$ by a one-parameter local group of diffeomorphisms, with the perturbation background η_{ab} being unchanged.

Suppose λ is so small that both the domain U and the range $\phi_\lambda[U]$ of the diffeomorphism $\phi_\lambda : U \rightarrow \phi_\lambda[U]$ are contained in the coordinate patch of $\{x^\mu\}$. Then four functions $y^\mu(\lambda) \equiv$

$\phi_{-\lambda}^* x^\mu$ ($\mu = 0, 1, 2, 3$) form a coordinate system on $\phi_\lambda[U]$. When $\lambda = \varepsilon$, the corresponding $y^\mu(\varepsilon)$ is denoted by x'^μ . Then, since ε is infinitesimal, we have

$$x'^\mu = y^\mu(\varepsilon) = \phi_{-\varepsilon}^* x^\mu = x^\mu - \varepsilon \mathcal{L}_X x^\mu = x^\mu - \mathcal{L}_\xi x^\mu = x^\mu - \xi^\mu,$$

where in the last step we used (4.2.2) and (2.2.3'), and ξ^μ are the components of ξ^a in $\{x^\mu\}$. This is exactly the infinitesimal coordinate transformation (7.8.19). Noticing that $g_{ab} = (\phi_\lambda)_* \tilde{g}_{ab}(\lambda)$, according to Theorem 4.1.3, $\forall q \in U$ the coordinate components of $g_{ab}|_{\phi_\lambda(q)}$ in $\{y^\mu(\lambda)\}$ equal the corresponding coordinate components of $\tilde{g}_{ab}(\lambda)|_q$ in $\{x^\mu\}$. Especially, for $\lambda = \varepsilon$, this yields

$$\tilde{g}_{cd}(\varepsilon)|_q \left(\frac{\partial}{\partial x^\mu} \right)^c \Big|_q \left(\frac{\partial}{\partial x^\nu} \right)^d \Big|_q = g_{cd}|_{\phi_\varepsilon(q)} \left(\frac{\partial}{\partial x'^\mu} \right)^c \Big|_{\phi_\varepsilon(q)} \left(\frac{\partial}{\partial x'^\nu} \right)^d \Big|_{\phi_\varepsilon(q)} = g'_{\mu\nu}(x')|_{\phi_\varepsilon(q)},$$

where $g'_{\mu\nu}(x')$ are the coordinate components of g_{cd} in $\{y^\mu(\varepsilon) \equiv x'^\mu\}$, as shown in (7.8.20). Hence

$$\tilde{g}_{ab}(\varepsilon)|_q = \tilde{g}_{cd}(\varepsilon)|_q \left(\frac{\partial}{\partial x^\mu} \right)^c \Big|_q \left(\frac{\partial}{\partial x^\nu} \right)^d \Big|_q (dx^\mu)_a|_q (dx^\nu)_b|_q = g'_{\mu\nu}(x')|_{\phi_\varepsilon(q)} (dx^\mu)_a|_q (dx^\nu)_b|_q.$$

Since $g'_{\mu\nu}(x')|_{\phi_\varepsilon(q)} = g'_{\mu\nu}(x'|_{\phi_\varepsilon(q)}) = g'_{\mu\nu}(x|_q)$, it turns out that $\tilde{g}_{ab}(\varepsilon) = g'_{\mu\nu}(x) (dx^\mu)_a (dx^\nu)_b$. Then, precisely to the order of ε , we have

$$\mathcal{L}_\xi g_{ab} = \varepsilon \mathcal{L}_X g_{ab} = \tilde{g}_{ab}(\varepsilon) - g_{ab} = [g'_{\mu\nu}(x) - g_{\mu\nu}(x)] (dx^\mu)_a (dx^\nu)_b.$$

This means that the coordinate components of $\mathcal{L}_\xi g_{ab}$ in $\{x^\mu\}$ are actually $g'_{\mu\nu}(x) - g_{\mu\nu}(x)$, not $g'_{\mu\nu}(x') - g_{\mu\nu}(x)$ in (7.8.21). However, their difference

$$\begin{aligned} & [g'_{\mu\nu}(x') - g_{\mu\nu}(x)] - [g'_{\mu\nu}(x) - g_{\mu\nu}(x)] = g'_{\mu\nu}(x') - g'_{\mu\nu}(x) \\ &= -\xi^\rho \frac{\partial g'_{\mu\nu}}{\partial x^\rho} + O(\varepsilon^2) = -\xi^\rho \frac{\partial g'_{\mu\nu}}{\partial x^\rho} + O(\varepsilon^2) = O(\varepsilon^2) \end{aligned}$$

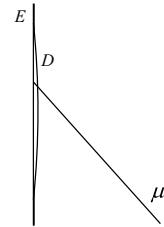
is negligible on the order of ξ^a . Therefore, the gauge transformation (7.8.14) and the infinitesimal coordinate transformation (7.8.19) are equivalent up to terms of higher order than ξ^a . In fact, what we just saw is a special case of gauge transformations in general relativity. We will come back to a more general discussion in Sect. 8.10.

[The End of Optional Reading 7.8.2]

7.8.2 The Newtonian Limit

In this subsection we will show that Newton's theory of gravity can be regarded as the limit of general relativity under the weak-field and low-speed condition. First, let us give an interpretation for the “weak-field and low-speed condition”. Take the gravitational field around the Earth as an example, it corresponds to a slightly curved metric field $g_{ab} = \eta_{ab} + \gamma_{ab}$, where γ_{ab} is “small”. In Fig. 7.14, E and D represent, respectively, the world lines of the Earth and a shell shot from a cannon on the ground (their relative speed $u_{ED} \ll 1$), and μ represents the world line of a “high speed” muon from a cosmic ray. Its “high speed” is from the perspective of an observer on the Earth; the muon regards itself as being at rest while E is moving at a high

Fig. 7.14 The world lines of the Earth E , the shell D and the muon



speed. Either way, their relative speed is close to the speed of light ($u_{\mu E} \cong 1$). As a flat metric field, η_{ab} has many inertial coordinate systems, such as the inertial frame $\{t, x^i\}$ which uses the world line of E as a t -coordinate line, and the inertial frame $\{t', x'^i\}$ which uses the world line of μ as a t' -coordinate line; these two systems differ by a boost. The 3-speeds of the Earth, the shell as well as cars, airplanes, etc. relative to the system $\{t, x^i\}$ are all very small, while the 3-speeds of them relative to $\{t', x'^i\}$ are large. The “weak-field and low-speed limit” should be interpreted as follows: there exists an inertial coordinate system of η_{ab} (in the example above it is $\{t, x^i\}$), in which all the objects we are concerned with have coordinate speeds much less than 1 (and thus in $\{t, x^i\}$ one cannot use the Newtonian theory to discuss a problem that involves a muon), and $|\gamma_{\mu\nu}| \equiv |g_{\mu\nu} - \eta_{\mu\nu}| \ll 1$.

Specifically speaking, the “weak-field and low-speed” condition guarantees that there exists an η_{ab} such that $\gamma_{ab} = g_{ab} - \eta_{ab}$ is “small”, and there exists an inertial coordinate system $\{t, x^i\}$ of η_{ab} which satisfies:

(1) The energy-momentum tensor T_{ab} of the source of the gravitational field can be expressed in this system as:

$$T_{ab} \cong \rho(dt)_a(dt)_b. \quad (7.8.22)$$

That is, only T_{00} , the time-time component of T_{ab} , is nonvanishing in this system. The space-time components T_{0i} vanish since the small velocity of the source leads to a small momentum density; the space-space components T_{ij} vanishing indicates that, compared with the mass density, the 3-dimensional stress can be ignored (for instance, the pressure p in the Earth’s center is only 10^{-10} times the density ρ). Thus, although in general relativity each component of the energy-momentum tensor T_{ab} of the matter field contributes to the spacetime curvature, in Newton’s theory of gravity (as is known to all) only the mass density ρ contributes to the gravitational field.

(2) (a) The spacetime geometry changes slowly due to the low-speed motion of the source, and hence $\partial\bar{\gamma}_{\mu\nu}/\partial t$ can be ignored; (b) the low-speed motion of an object in the gravitational field leads to the fact that its 4-velocity U^a is approximately equal to the 4-velocity $Z^a \equiv (\partial/\partial t)^a$ of an observer in the $\{t, x^i\}$ system, i.e., $U^a \cong Z^a$.

The linearized Einstein equation under the Lorenz gauge condition can be simplified under the above approximations:

$$\text{components of the l.h.s. of (7.8.16)} = \partial^\sigma \partial_\sigma \bar{\gamma}_{\mu\nu} = \partial^0 \partial_0 \bar{\gamma}_{\mu\nu} + \partial^i \partial_i \bar{\gamma}_{\mu\nu} \cong \partial^i \partial_i \bar{\gamma}_{\mu\nu} = \nabla^2 \bar{\gamma}_{\mu\nu},$$

where we used the approximation condition (2) in the third equality, and ∇^2 is the square of the derivative operator $\bar{\nabla}$ in the 3-dimensional coordinate system $\{x^i\}$. On the other hand, from the approximate condition (1) we can see that the components of the right-hand side of (7.8.16) $\cong -16\pi\rho$ when $\mu = \nu = 0$, and the other components vanish, i.e.,

$$\nabla^2 \bar{\gamma}_{00} = -16\pi\rho, \quad (7.8.23)$$

$$\nabla^2 \bar{\gamma}_{0i} = 0, \quad (7.8.24)$$

$$\nabla^2 \bar{\gamma}_{ij} = 0. \quad (7.8.24')$$

The unique solutions $\bar{\gamma}_{0i}$ and $\bar{\gamma}_{ij}$ for equations (7.8.24) and (7.8.24') that are well-behaved at infinity are constants, which can be set to zero by means of a gauge transformation. Thus, the only nonzero component of $\bar{\gamma}_{\mu\nu}$ is $\bar{\gamma}_{00}$, which satisfies equation (7.8.23). Let

$$\phi \equiv -\frac{1}{4}\bar{\gamma}_{00}, \quad (7.8.25)$$

and interpret ϕ as the Newtonian gravitational potential, then equation (7.8.23) will become the well-known Poisson equation in Newton's theory of gravity:

$$\nabla^2 \phi = 4\pi\rho. \quad (7.8.26)$$

The conclusion that the only nonzero component of $\bar{\gamma}_{\mu\nu}$ is $\bar{\gamma}_{00}$ can also be expressed in terms of a tensor equation as

$$\bar{\gamma}_{ab} = \bar{\gamma}_{00}(dt)_a(dt)_b = -4\phi(dt)_a(dt)_b. \quad (7.8.27)$$

Hence,

$$\bar{\gamma} \equiv \eta^{ab}\bar{\gamma}_{ab} = \bar{\gamma}_{00}\eta^{ab}(dt)_a(dt)_b = -\bar{\gamma}_{00} = 4\phi. \quad (7.8.28)$$

Also, from $\gamma_{ab} = \bar{\gamma}_{ab} + \eta_{ab}\gamma/2$ we get $\gamma = \eta^{ab}\gamma_{ab} = \eta^{ab}\bar{\gamma}_{ab} + \eta^{ab}\eta_{ab}\gamma/2 = \bar{\gamma} + 2\gamma$, and thus $\gamma = -\bar{\gamma}$. Therefore,

$$\gamma_{ab} = \bar{\gamma}_{ab} - \frac{1}{2}\eta_{ab}\bar{\gamma}. \quad (7.8.29)$$

By means of (7.8.27) and (7.8.28), the above equation can be rewritten as

$$\gamma_{ab} = -\phi[4(dt)_a(dt)_b + 2\eta_{ab}]. \quad (7.8.30)$$

Based on the discussion above, we can derive the equation of motion for a point mass under the Newtonian approximation. Suppose there is no force acting on the point mass other than gravity, then from the viewpoint of general relativity its world line should be a geodesic, whose equation in the inertial coordinate system of η_{ab} is

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\nu\sigma} \frac{dx^\nu}{d\tau} \frac{dx^\sigma}{d\tau} = 0, \quad (7.8.31)$$

where τ is the proper time of the point mass. Under the Newtonian approximation, the condition $U^a \cong Z^a$ satisfied by the 4-velocity U^a of the point mass assures that $\tau \cong t$ (the proper time is approximately equal to the coordinate time) and $u^i \equiv dx^i/dt \cong 0$ (the 3-velocity is approximately zero), and hence $U^\nu \equiv dx^\nu/dt$ is approximately $(1, 0, 0, 0)$. Therefore, (7.8.31) can be expressed approximately as

$$\frac{d^2x^\mu}{dt^2} = -\Gamma^\mu_{00}. \quad (7.8.32)$$

It follows from (7.8.4) that [the superscript (1) of Γ is omitted]

$$\begin{aligned} \Gamma^0_{00} &= \frac{1}{2}\eta^{00}(\gamma_{00,0} + \gamma_{00,0} - \gamma_{00,0}) = -\frac{1}{2}\frac{\partial\gamma_{00}}{\partial t} \cong 0, \\ \Gamma^i_{00} &= \frac{1}{2}\eta^{ij}(\gamma_{j0,0} + \gamma_{0j,0} - \gamma_{00,j}) \cong -\frac{1}{2}\delta^{ij}\gamma_{00,j} = -\frac{1}{2}\frac{\partial\gamma_{00}}{\partial x^i}, \quad i = 1, 2, 3, \end{aligned} \quad (7.8.33)$$

where the second equality is due to $\gamma_{j0} = \bar{\gamma}_{j0} + \frac{1}{2}\gamma\eta_{j0} = 0$. Hence, (7.8.32) gives an identity when $\mu = 0$, and gives $\frac{d^2x^i}{dt^2} = \frac{1}{2}\frac{\partial\gamma_{00}}{\partial x^i}$ ($i = 1, 2, 3$) when $\mu = i$. Then, from (7.8.29) and (7.8.28) we have $\gamma_{00} = \bar{\gamma}_{00}/2 = -2\phi$. Plugging this into the equation above, and noticing (7.8.25), we obtain $d^2x^i/dt^2 = -\partial\phi/\partial x^i$. Since d^2x^i/dt^2 is the i th component of the 3-acceleration \vec{a} of the point mass relative to the inertial coordinate system $\{t, x^i\}$, the above equation can be expressed in terms of an equality of 3-vectors as

$$\vec{a} = -\vec{\nabla}\phi. \quad (7.8.34)$$

This is exactly the equation of motion for a point mass that only undergoes gravitational force in Newton's theory of gravity. Equations (7.8.26) and (7.8.34) are the basic equations of Newton's theory of gravity; thus, Newton's theory of gravity can be regarded as the weak-field and low-speed limit of general relativity. From

$$\phi \equiv -\frac{1}{4}\bar{\gamma}_{00} = -\frac{1}{2}\gamma_{00} \quad (7.8.35)$$

we have $g_{00} = \eta_{00} + \gamma_{00} = -(1 + 2\phi)$, or,

$$\phi = -\frac{1}{2}(1 + g_{00}). \quad (7.8.36)$$

This reflects the close relation between the metric component g_{00} and the Newtonian gravitational potential under the Newtonian approximation. Again, take the balls 1 and 2 in Fig. 7.10 as an example. Choose the inertial coordinate system $\{t, x, y, z\}$ of η_{ab} such that the z -axis is vertically upwards, then $Z^a = (\partial/\partial t)^a$ in (7.6.8).

Noticing (7.6.7), one can rewrite (7.6.8) as $\tilde{a}^c = -R_{0b0}{}^c \lambda^b$, whose z -component is $\tilde{a}^z = -R_{0z0z} \lambda^z$ (z is not summed over). In the Newtonian approximation, the derivative with respect to t can be ignored, and hence it follows from (7.8.5) that $R_{0z0z} = -\frac{1}{2} \partial^2 \gamma_{00} / \partial z^2 = \partial^2 \phi / \partial z^2 = d^2 \phi / dr^2$, and therefore $\tilde{a}^z = -(d^2 \phi / dr^2) \lambda^z$, which is in agreement with (7.6.2). This is a verification of the following statement: the tidal acceleration in general relativity determined by the curvature tensor according to (7.6.8) will return to the tidal acceleration in the Newtonian mechanics determined by (7.6.2) under the weak-field and low-speed approximation.

7.9 Gravitational Radiation

The resemblance between the gravitational field and the electromagnetic field makes people expect that there exists gravitational radiation in general relativity similar to the electromagnetic radiation. Actually, the fact that there exists a wave solution to Einstein's equation which propagates at the speed of light was already well-known soon after general relativity was published. Nevertheless, for quite a while the authenticity of gravitational waves was in doubt. A. S. Eddington suggested in 1922 that a gravitational wave solution only represents the wave motion of the spacetime coordinates, and thus has no observational effect. The situation has turned around since the 1950s. Using a coordinate independent method, H. Bondi and collaborators showed that gravitational waves indeed carry energy and momentum, and the mass of the system must decrease when it emits gravitational waves. This led to the physical authenticity and observability of gravitational radiation being gradually accepted.

7.9.1 Gauge Conditions of the Linearized Theory of Gravity

First we will discuss the gravitational waves under the approximation of linearized gravity. Before introducing the wave solutions to the linearized Einstein equation, let us first discuss some useful gauge conditions in the linearized theory of gravity.

As we have seen in Sect. 7.8.1, the Lorenz gauge condition

$$\partial^b \bar{\gamma}_{ab} = \partial^b \gamma_{ab} - \frac{1}{2} \partial_a \gamma = 0 \quad (7.9.1)$$

in linearized gravity is inspired by the Lorenz gauge condition $\partial^a A_a = 0$ of the electromagnetic field. However, in electrodynamics, $\partial^a A_a = 0$ and the wave equation (with source) $\partial^b \partial_b A^a = -4\pi J^a$ cannot determine the 4-potential A_a completely, since another 4-potential

$$A'_a = A_a + \partial_a \chi \quad (7.9.2)$$

also satisfies $\partial^a A'_a = 0$ and $\partial^a \partial_a A'_b = -4\pi J_b$ as long as

$$\partial^a \partial_a \chi = 0. \quad (7.9.3)$$

In a source-free ($J^a = 0$) region, it can be proved that there exists a function χ satisfying the above condition such that $A'_0 = 0$. The condition $A'_0 = 0$ together with the Lorenz gauge condition $\partial^a A'_a = 0$ is called the **radiation gauge condition**. In order to find such a χ , we first find a function χ' satisfying

$$\frac{\partial \chi'}{\partial t} = -A_0. \quad (7.9.4)$$

As long as A_0 is a C^2 function, there is no problem for the existence of χ' . Then, in a source-free region we have

$$\frac{\partial}{\partial t}(\partial^a \partial_a \chi') = \partial^a \partial_a \frac{\partial \chi'}{\partial t} = -\partial^a \partial_a A_0 = 0,$$

which indicates that $\partial^a \partial_a \chi'$ is time-independent. Then, we can find a time-independent function χ_0 that satisfies

$$\nabla^2 \chi_0 = \partial^a \partial_a \chi'. \quad (7.9.5)$$

It is easy to verify that $\chi = \chi' - \chi_0$ is the function χ that makes $A'_a = A_a + \partial_a \chi$ satisfy the radiation gauge condition. To put it more precisely, we have the following proposition:

Proposition 7.9.1 *Let U be a non-empty open subset of a spacetime, on which a flat Lorentzian metric η_{ab} is defined, and let $\{x^\mu\}$ ($x^0 \equiv t$) be a coordinate system where the components of η_{ab} are $\eta_{\mu\nu}$.¹⁸ Suppose A_0 is a C^2 function on U satisfying $\partial^a \partial_a A_0 = 0$. Then, $\forall p \in U$ there exists a C^3 function χ on U satisfying $\partial^a \partial_a \chi = 0$ and $\partial \chi / \partial t = -A_0$ in an open neighborhood $U' \subset U$ of p . Moreover, if A_0 is smooth, then there exists a smooth χ on U satisfying $\partial^a \partial_a \chi = 0$ and $\partial \chi / \partial t = -A_0$ on U' .*

Proof [Optional Reading]

According to the discussion above, all we have to prove is that there exists a χ_0 satisfying (7.9.5) in some neighborhood U' of p . The Lorentzian system $\{x^\mu\}$ defines a chart (U, ψ) of M . Then one can choose a neighborhood $U' \subsetneq U$ of p such that $\psi[U'] = I \times \Sigma' \subset \mathbb{R}^4$, where I is an open interval and $\Sigma' \subset \mathbb{R}^3$ is enclosed by a closed piecewise smooth surface. In \mathbb{R}^3 , one can choose an open ball Σ satisfying $\Sigma' \subsetneq \Sigma$. Then, one can construct a time-independent C^1 (or smooth if so is A_0) function ρ on \mathbb{R}^4 as follows:

¹⁸ Here we only require η_{ab} to be flat on U , and the same for Proposition 7.9.2 (see the first footnote in Sect. 7.8). It follows from Theorem 3.4.9 that for any (locally) flat metric there exists a coordinate system such that the metric components are constant. For Lorentzian signature, one can further find a coordinate transformation and turn them into $\eta_{\mu\nu}$.

$$\rho(\vec{x}) = \begin{cases} (\partial^a \partial_a \chi')|_q, & \text{if } (t, \vec{x}) = \psi(q) \in \psi[U'], \\ 0, & \text{if } (t, \vec{x}) \in \mathbb{R}^4 - \psi^{-1}[\mathbb{R} \times \Sigma]. \end{cases}$$

Then, there exists the following integral:

$$\phi(\vec{x}) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{\rho(\vec{x}')}{|\vec{x} - \vec{x}'|} dx'^1 \wedge dx'^2 \wedge dx'^3 = -\frac{1}{4\pi} \int_{\Sigma} \frac{\rho(\vec{x}')}{|\vec{x} - \vec{x}'|} dx'^1 \wedge dx'^2 \wedge dx'^3,$$

which satisfies Poisson's equation $\nabla^2 \phi = \rho$ (this is a well-known result in electrostatics). It is easy to see that ϕ is C^3 if A_0 is C^2 and ϕ is smooth if A_0 is smooth. Now define $\chi_0 = \psi^* \phi$, then $\nabla^2 \chi_0 = \psi^* \rho$, which equals $\partial_a \partial^a \chi'$ when restricted to U' . \square

The situation of linearized gravity is very similar: the linearized Einstein equation and the Lorenz gauge condition $\partial^a \bar{\gamma}_{ab} = 0$ cannot determine γ_{ab} completely, since if we set

$$\gamma'_{ab} = \gamma_{ab} + \partial_a \xi_b + \partial_b \xi_a, \quad (7.9.6)$$

then γ'_{ab} also satisfies (7.8.16) and $\partial^a \bar{\gamma}'_{ab} = 0$ as long as ξ_a satisfies

$$\partial^b \partial_b \xi_a = 0. \quad (7.9.7)$$

(The existence of such a ξ^a will be proved in Optional Reading 7.9.1). In a source-free region, one can further set $\gamma = 0$ and $\gamma_{0i} = 0$ ($i = 1, 2, 3$). Together with the Lorenz gauge condition, this is called the **radiation gauge condition** of the linearized theory of gravity. Furthermore, as we will show below, one can also set $\gamma_{00} = 0$, and the gauge condition becomes

$$\partial^b \bar{\gamma}_{ab} = 0, \quad \gamma = 0, \quad \gamma_{0v} = 0, \quad v = 0, 1, 2, 3, \quad (7.9.8)$$

called the **transverse-traceless gauge condition**, or TT gauge condition for short.

Proposition 7.9.2 *Let U be a non-empty open subset of a spacetime, on which a flat Lorentzian metric η_{ab} is defined, and let $\{x^\mu\}$ ($x^0 \equiv t$) be a coordinate system where the components of η_{ab} are $\eta_{\mu\nu}$. Suppose γ_{ab} is a smooth symmetric tensor field which satisfies on U the Lorenz gauge condition $\partial^a \bar{\gamma}_{ab} = 0$ and*

$$\partial^c \partial_c \gamma = 0, \quad \partial^c \partial_c \gamma_{0v} = 0, \quad v = 0, 1, 2, 3. \quad (7.9.9)$$

Then, $\forall p \in U$ there exists a smooth vector field ξ^a on U and an open neighborhood $U' \subset U$ of p such that $\gamma'_{ab} = \gamma_{ab} + \partial_a \xi_b + \partial_b \xi_a$ satisfies the transverse-traceless gauge condition on U' .

Proof See Optional Reading 7.9.1. \square

Note that in the above proposition, γ_{ab} is not necessarily a solution to the linearized Einstein equation. Now we consider γ_{ab} as a solution to the source-free linearized Einstein equation in the Lorenz gauge, then (7.8.16) with $T_{ab} = 0$ is reduced to

$$\partial^c \partial_c \gamma_{ab} - \frac{1}{2} \eta_{ab} \partial^c \partial_c \gamma = 0. \quad (7.9.10)$$

Contracting both sides of the above equation with η^{ab} yields $\partial^c \partial_c \gamma = 0$, and (7.9.10) becomes

$$\partial^c \partial_c \gamma_{ab} = 0. \quad (7.9.11)$$

In this case, the conditions in (7.9.9) are both satisfied. In fact, it is obvious that (7.9.10) and (7.9.11) are equivalent to each other, since if one is satisfied, so is the other. From the Lorenz gauge condition (7.9.1) we also see that $\partial^a \partial^b \gamma_{ab} = \frac{1}{2} \partial^a \partial_a \gamma$, and hence $\partial^c \partial_c \gamma = 0$ also leads to

$$\partial^a \partial^b \gamma_{ab} = 0. \quad (7.9.12)$$

As we have discussed above, given a solution γ_{ab} of the source-free linearized Einstein equation satisfying the Lorenz gauge condition, $\gamma'_{ab} = \gamma_{ab} + \partial_a \xi_b + \partial_b \xi_a$ is automatically a solution of the same equation as long as it also satisfies the Lorenz gauge condition ($\partial^b \partial_b \xi_a = 0$). Applying this to Proposition 7.9.2, we have the following conclusion:

Corollary 7.9.3 *Suppose a smooth symmetric tensor field γ_{ab} is a solution of the the source-free linearized Einstein equation satisfying the Lorenz gauge condition. Then, for each point p in the domain U of γ_{ab} , there exists $\gamma'_{ab} = \gamma_{ab} + \partial_a \xi_b + \partial_b \xi_a$ in an open neighborhood $U' \subset U$ of p , which is a solution of the source-free linearized Einstein equation satisfying the transverse-traceless gauge condition.*

Now let us count the degrees of freedoms of γ_{ab} in the TT gauge. It follows from $\gamma_{\mu\nu} = \gamma_{\nu\mu}$ that γ_{ab} has at most 10 independent components, while they are also constrained by (7.9.8). The conditions in (7.9.8) contain in total $4 + 4 + 1 = 9$ equations, but $\partial^\nu \gamma_{0\nu} = 0$ is also an outcome of $\gamma_{0\nu} = 0$, and so among these 9 equations only 8 are independent. Therefore, γ_{ab} has only $10 - 8 = 2$ independent components.¹⁹ Later we will see that in physics they correspond to the two independent polarization states (modes) of gravitational plane waves, see Sect. 7.9.2.

For the linearized Einstein equation (not necessarily source-free), there are also some other common gauge conditions, such as the **transverse gauge condition**, which requires

¹⁹ Note that this is a handwaving discussion, since the constraint counting is actually very subtle when it comes to partial differential equations. For example, the second equation in (7.9.15) can be regarded as a constraint for ξ_0 in the first equation, but it does not mean that ξ_0 has no degree of freedom! For another example, the 1-dimensional wave equation $\partial_t^2 u - c^2 \partial_x^2 u = 0$ has the general solution $u = f_+(x - ct) + f_-(x + ct)$, with f_\pm being arbitrary C^2 functions of one variable. If the wave equation is considered to be a constraint, is the number of constraints 1 or -1 ?

$$\partial_i \gamma^{0i} = 0, \quad \partial_i s^{ij} = 0 \quad (\text{where } s_{ij} = \gamma_{ij} - \frac{1}{3} \delta^{kl} \gamma_{kl} \delta_{ij}), \quad i, j = 1, 2, 3, \quad (7.9.13)$$

and the **synchronous gauge condition**, which requires

$$\gamma_{0\mu} = 0, \quad \mu = 0, 1, 2, 3. \quad (7.9.14)$$

The reader may refer to Carroll (2019) for more discussions about these gauge conditions.

[Optional Reading 7.9.1]

Proof of Proposition 7.9.2 (1) According to Proposition 7.9.1, there exists a function ξ_0 on U such that $\forall p \in U$,

$$\partial_c \partial^c \xi_0 = 0, \quad \frac{\partial \xi_0}{\partial t} = -\frac{1}{2} \gamma_{00} \quad (7.9.15)$$

are both satisfied on a neighborhood U'_0 of p .

(2) For each of $i = 1, 2, 3$, there is obviously a smooth function ξ'_i on U satisfying

$$\frac{\partial \xi'_i}{\partial t} = -\gamma_{0i} - \frac{\partial \xi_0}{\partial x^i}. \quad (7.9.16)$$

Then, using (7.9.16) and the second equation of (7.9.15) we can derive that

$$\partial_c \partial^c \xi'_i = -\frac{\partial^2 \xi'_i}{\partial t^2} + \nabla^2 \xi'_i = \frac{\partial \gamma_{0i}}{\partial t} + \frac{\partial}{\partial x^i} \frac{\partial \xi_0}{\partial t} + \nabla^2 \xi'_i = \frac{\partial \gamma_{0i}}{\partial t} - \frac{1}{2} \frac{\partial \gamma_{00}}{\partial x^i} + \nabla^2 \xi'_i. \quad (7.9.17)$$

On the other hand, from (7.9.16) we also have on U'_0 that

$$\frac{\partial}{\partial t} \partial_c \partial^c \xi'_i = \partial_c \partial^c \frac{\partial \xi'_i}{\partial t} = -\partial_c \partial^c \gamma_{0i} - \frac{\partial}{\partial x^i} \partial_c \partial^c \xi_0 = 0,$$

where (7.9.9) and the first equation in (7.9.15) are used in the last step. Thus, the right side of (7.9.17) is independent of t on U'_0 , and thus there exist smooth functions X'_i ($i = 1, 2, 3$) that satisfy on U'_0

$$\frac{\partial X'_i}{\partial t} = 0, \quad \nabla^2 X'_i = -\frac{\partial \gamma_{0i}}{\partial t} + \frac{1}{2} \frac{\partial \gamma_{00}}{\partial x^i} - \nabla^2 \xi'_i. \quad (7.9.18)$$

Combining (7.9.16), (7.9.17) and (7.9.18), we can see that each function $\xi'_i + X'_i$ on U'_0 satisfies

$$\frac{\partial(\xi'_i + X'_i)}{\partial t} + \frac{\partial \xi_0}{\partial x^i} = -\gamma_{0i}, \quad \partial_c \partial^c(\xi'_i + X'_i) = 0. \quad (7.9.19)$$

(3) For convenience, denote $\vec{\xi}' \equiv (\xi'_1, \xi'_2, \xi'_3)$ and $\vec{X}' \equiv (X'_1, X'_2, X'_3)$. For example, the notation $\vec{\nabla} \cdot \vec{\xi}'$ can be regarded as an abbreviation of $\delta^{ij} \frac{\partial \xi'_j}{\partial x^i}$. From the first equation in (7.9.19), we obtain on U'_0

$$\vec{\nabla} \cdot \frac{\partial \vec{\xi}'}{\partial t} + \vec{\nabla} \cdot \frac{\partial \vec{X}'}{\partial t} + \nabla^2 \xi_0 = -\delta^{ij} \frac{\partial \gamma_{0j}}{\partial x^i}. \quad (7.9.20)$$

Then, one can find on U'_0 that

$$\frac{\partial}{\partial t} \left(-\frac{1}{2}(\gamma_{00} + \gamma) - \vec{\nabla} \cdot \vec{\xi}' - \vec{\nabla} \cdot \vec{X}' \right) = 0.$$

[The reader should complete the proof. Hint: use (7.9.20), (7.9.1) and (7.9.15)]. Thus, $-\frac{1}{2}(\gamma_{00} + \gamma) - \vec{\nabla} \cdot \vec{\xi}' - \vec{\nabla} \cdot \vec{X}'$ is independent of t when restricted to U'_0 . This allows us to find a function ϕ defined on an open neighborhood $U'_\phi \subset U'_0$ of p such that

$$\frac{\partial \phi}{\partial t} = 0, \quad \nabla^2 \phi = \frac{1}{2}(\gamma_{00} + \gamma) + \vec{\nabla} \cdot \vec{\xi}' + \vec{\nabla} \cdot \vec{X}'. \quad (7.9.21)$$

(4) Applying ∇^2 on both sides of the second equation in (7.9.21), one finds on U'_ϕ that

$$\nabla^2 \nabla^2 \phi = 0. \quad (7.9.22)$$

[The reader should complete the proof. Hint: use (7.9.18), (7.9.1) and (7.9.9)]. This is equivalent to say that $\vec{\nabla} \cdot \vec{\nabla} \nabla^2 \phi = 0$, namely the 3-vector field $\vec{\nabla} \nabla^2 \phi$ is divergence-free. Thus, there exists an open neighborhood $U''_\phi \subset U'_\phi$ of p diffeomorphic to \mathbb{R}^4 such that $\vec{\nabla} \nabla^2 \phi = \vec{\nabla} \times \vec{Y}$ is satisfied on U''_ϕ for some 3-vector field \vec{Y} defined on U . Since ϕ does not depend on t when restricted on U'_ϕ , we can require that \vec{Y} does not depend on t on U''_ϕ . Thus, there exists a 3-vector field \vec{X} on U which is independent of t such that $\nabla^2 \vec{X} = \vec{Y}$ is satisfied on an open neighborhood $U' \subset U''_\phi$ of p . Then, we have on U' that

$$\nabla^2 (\vec{\nabla} \times \vec{X} - \vec{\nabla} \phi) = 0. \quad (7.9.23)$$

(5) So far we have introduced a series of functions and 3-vector fields, whose domains are open neighborhoods of p . Since we do not care about their behaviors outside these neighborhoods, they can be extended arbitrarily to smooth functions or 3-vector fields on U . Thus, from now on, all concerned functions and 3-vector fields are defined on U , while the equations they satisfy are valid on $U' \subset U$.

Now we define a 3-vector field $\vec{\xi} = (\xi_1, \xi_2, \xi_3)$ on U as follows:

$$\vec{\xi} = \vec{\xi}' + \vec{X}' - \vec{\nabla} \phi + \vec{\nabla} \times \vec{X}. \quad (7.9.24)$$

When restricted to $U' \subset U''_\phi \subset U'_\phi \subset U'_0 \subset U$, both ϕ and \vec{X} are independent of t , and so (7.9.19) gives

$$\frac{\partial \xi_i}{\partial t} + \frac{\partial \xi_0}{\partial x^i} = -\gamma_{0i}, \quad (7.9.25)$$

$$\partial_c \partial^c \vec{\xi} = \partial_c \partial^c (\vec{\nabla} \times \vec{X} - \vec{\nabla} \phi) = \nabla^2 (\vec{\nabla} \times \vec{X} - \vec{\nabla} \phi) = 0, \quad (7.9.26)$$

where (7.9.23) is used in the last step of (7.9.26). Using the second equation in (7.9.21), we have

$$\vec{\nabla} \cdot \vec{\xi} = \vec{\nabla} \cdot \vec{\xi}' + \vec{\nabla} \cdot \vec{X}' - \nabla^2 \phi = -\frac{1}{2}(\gamma_{00} + \gamma).$$

Now, let ξ_0, ξ_1, ξ_2 and ξ_3 be the coordinate components of a 1-form ξ_a on U . Then, combining the above equation and the second equation in (7.9.15) yields

$$\partial_a \xi^a = \eta^{\mu\nu} \frac{\partial \xi_\nu}{\partial x^\mu} = -\frac{\partial \xi_0}{\partial t} + \vec{\nabla} \cdot \vec{\xi} = -\frac{1}{2}\gamma. \quad (7.9.27)$$

Similarly, the wave equations (7.9.15) and (7.9.26) for ξ_0 and $\vec{\xi}$ can be combined into

$$\partial_c \partial^c \xi_a = 0. \quad (7.9.28)$$

Finally, the second equation in (7.9.15) can be combined with (7.9.25) into

$$\frac{\partial \xi_\nu}{\partial t} + \frac{\partial \xi_0}{\partial x^\nu} = -\gamma_{0\nu}. \quad (7.9.29)$$

(6) Now let us consider the tensor field $\gamma'_{ab} = \gamma_{ab} + \partial_a \xi_b + \partial_b \xi_a$ on U . It follows that $\gamma' = \eta^{ab} \gamma'_{ab} = \gamma + 2 \partial_a \xi^a$. From now on the equations will be restricted on U' . First, we can see that $\gamma' = 0$ due to (7.9.27). From (7.9.27) and (7.9.28) we obtain that $\partial^b \gamma'_{ab} = \partial^b \gamma_{ab} - \frac{1}{2} \partial_a \gamma = 0$, i.e., γ'_{ab} satisfies the Lorenz gauge condition. Also, it follows from (7.9.29) that

$$\gamma'_{0\nu} = \gamma_{0\nu} + \frac{\partial \xi_\nu}{\partial t} + \frac{\partial \xi_0}{\partial x^\nu} = 0, \quad \nu = 0, 1, 2, 3.$$

Having these, we have proved the existence of a gauge transformation such that γ'_{ab} satisfies the TT gauge condition on U' . \square

[The End of Optional Reading 7.9.1]

7.9.2 Gravitational Plane Waves

The source-free linearized Einstein equation is a good description for the gravitational waves emitted by a source far away from an observer. In Sect. 7.8, we have seen that under a gauge transformation γ_{ab} satisfies the Lorenz gauge condition. Then, according to Corollary 7.9.3, a further gauge transformation can make it satisfy the transverse-traceless (TT) gauge condition at least in an open neighborhood of the observer. Under the TT gauge condition, now we will investigate wave solutions of the source-free linearized Einstein equation.

In the TT gauge, γ_{ab} satisfies (7.9.8). The traceless condition reduces the Lorenz gauge condition to $\partial^b \gamma_{ab} = 0$, and the source-free linearized Einstein equation becomes (7.9.11). From now on, all the equations in this subsection are valid on an open neighborhood U of the observer's world line, on which a flat Lorentzian metric η_{ab} is defined, whose components in a coordinate system $\{x^\mu\}$ are $\eta_{\mu\nu}$. Then, the ordinary derivative operator ∂_a of the coordinate system $\{x^\mu\}$ is the derivative operator associated with η_{ab} .

As an ansatz, let us consider a solution to (7.9.11) of the following form:

$$\gamma_{ab} = f(K_\mu x^\mu) H_{ab}, \quad (7.9.30)$$

where f is a C^2 function of one variable, $K^\nu = K_\mu \eta^{\mu\nu}$ are the components of a constant 4-vector field K^a in $\{x^\mu\}$, and H_{ab} is a constant symmetric tensor field of type $(0, 2)$. K^a and H_{ab} being constant vector and tensor fields, respectively, means that

$$\partial_b K^a = 0, \quad \partial_c H_{ab} = 0. \quad (7.9.31)$$

In other words, all the components of K^a and H_{ab} in $\{x^\mu\}$ are constants. Note that γ_{ab} in (7.9.30) remains unchanged if we replace f by Cf and H_{ab} by H_{ab}/C for any nonzero constant C . Hence, if the range of f is bounded, we can assume that $-1 \leq f \leq 1$ with $|f(K_\mu x^\mu)| = 1$ at some spacetime point. In this way, H_{ab} represents the amplitude of the wave solution (7.9.30), called the **polarization tensor**, and K^a will be the wave 4-vector for a gravitational wave.

Noticing that

$$\partial_c (K_\mu x^\mu) = K_\mu \partial_c x^\mu = K_\mu (\mathrm{d}x^\mu)_c = K_c, \quad (7.9.32)$$

we have

$$\partial_c \gamma_{ab} = f'(K_\mu x^\mu) K_c H_{ab}, \quad \partial_c \partial_d \gamma_{ab} = f''(K_\mu x^\mu) K_c K_d H_{ab}, \quad (7.9.33)$$

where f' and f'' are the first and the second order derivatives of f , respectively. Hence, to obtain a solution that is nonzero and non-constant, we should consider $f \neq 0$, $f' \neq 0$ and $H_{ab} \neq 0$ (meaning that they are not identically zero, with possible zero points). Then, the TT gauge condition is now equivalent to

$$K^b H_{ab} = 0, \quad \eta^{ab} H_{ab} = 0, \quad H_{0v} = H_{v0} = 0, \quad v = 0, 1, 2, 3. \quad (7.9.34)$$

Plugging (7.9.30) into (7.9.11) yields

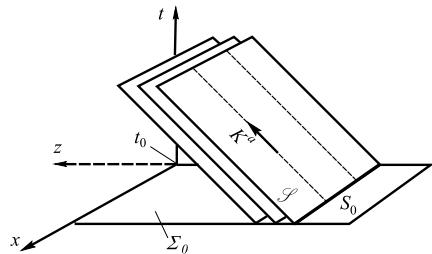
$$K^c K_c H_{ab} f''(K_\mu x^\mu) = 0. \quad (7.9.35)$$

First we consider a special case, namely $f'' = 0$. Since we have assumed that $f' \neq 0$, without loss of generality, we can set f to be $f(\lambda) = \lambda + \lambda_0$ with λ_0 a constant. The corresponding γ_{ab} then reads

$$\gamma_{ab} = (K_\mu x^\mu + \lambda_0) H_{ab}. \quad (7.9.36)$$

Since $\partial_c \partial_d \gamma_{ab} = 0$, it follows from (7.8.5) that the first-order Riemann curvature of $g_{ab} = \eta_{ab} + \gamma_{ab}$ vanishes, i.e., g_{ab} is a flat metric in the linear (first-order) approximation. Then, there exists a coordinate system $\{x'^\mu\}$ such that $g_{ab} = \eta_{\mu\nu} (\mathrm{d}x'^\mu)_a (\mathrm{d}x'^\nu)_b$ (see the first footnote in Sect. 7.9.1) in the linear approximation. (Note that the above coordinate transformation does not correspond to a gauge transformation described in Optional Reading 7.8.1). Therefore, a solution of the form (7.9.36) turns out to be

Fig. 7.15 The spacetime diagram of the gravitational plane wave propagating along the z -axis. \mathcal{S} is a constant-phase surface in the spacetime, and S_0 is the wavefront at a time t_0



a trivial solution at least in the linear approximation, and hence it is not regarded as having any physical effect.

From now on we will assume $f'' \neq 0$. In this case, (7.9.35) implies

$$K^a K_a = 0. \quad (7.9.37)$$

If $K^a = 0$, the corresponding γ_{ab} is a constant tensor field, which is not interesting in physics. Thus, we will focus on the case of $K^a \neq 0$, i.e., K^a is a nonzero null vector. Under the ansatz (7.9.30) with $f'' \neq 0$, (7.9.34) and (7.9.37) are the necessary and sufficient condition to determine a nontrivial solution of the source-free linearized Einstein equation in the TT gauge. Such a solution represents a **(traveling) gravitational plane wave**, whose wavefronts are surfaces described by $\Phi(x) \equiv K_\mu x^\mu = 0$. As shown in (7.9.32), $K_a = (d\Phi)_a$ is the normal covector of each of surfaces. It follows from (7.9.37) that

$$g^{ab} K_a K_b = (\eta^{ab} - \gamma^{ab}) K_a K_b = \eta^{ab} K_a K_b - f(K_\mu x^\mu) H^{ab} K_a K_b = \eta^{ab} K_a K_b = 0,$$

where (7.9.34) is used in the third equality. This indicates that in the linear approximation, a wavefront \mathcal{S} is a null surface with respect to either η_{ab} or g_{ab} . In other words, gravitational waves described by (7.9.30) propagate at the speed of light in vacuum just in a way similar to electromagnetic waves in Sect. 6.6.5, see Fig. 7.15. ($g_{ab} = \eta_{ab} + \gamma_{ab}$ and its curvature $R^a{}_{bcd}$ correspond to the electromagnetic 4-potential A_a and the electromagnetic field F_{ab} , respectively).

To demonstrate an important property of this plane wave solution, let us define $\tilde{K}^a = g^{ab} K_b$ (note that in linearized gravity we stipulate that $K^a = \eta^{ab} K_b$). In the linear approximation we have $g^{ab} = \eta^{ab} - f(K_\mu x^\mu) H^{ab}$, and thus up to higher order terms,

$$\tilde{K}^a = \eta^{ab} K_b - f(K_\mu x^\mu) H^{ab} K_b = K^a,$$

where (7.9.34) is used in the second equality. By means of the linearity and the Leibniz rule of $\mathcal{L}_{\tilde{K}}$, the Lie derivative of $g_{ab} = \eta_{ab} + f(K_\mu x^\mu) H_{ab}$ can be written as

$$\mathcal{L}_{\tilde{K}} g_{ab} = \mathcal{L}_K \eta_{ab} + H_{ab} \mathcal{L}_K [f(K_\nu x^\nu)] + f(K_\nu x^\nu) \mathcal{L}_K H_{ab}.$$

Using (4.2.8) and setting the ∇_a therein to the ordinary derivative ∂_a in $\{x^\mu\}$, we have

$$\mathcal{L}_K H_{ab} = K^c \partial_c H_{ab} + H_{cb} \partial_a K^c + H_{ac} \partial_b K^c = 0,$$

where (7.9.31) is used in the last step. Similarly one finds $\mathcal{L}_K \eta_{ab} = 0$. Then,

$$\mathcal{L}_{\tilde{K}} g_{ab} = H_{ab} \mathcal{L}_K [f(K_\nu x^\nu)] = H_{ab} K^c \partial_c [f(K_\nu x^\nu)] = H_{ab} K^c K_c f'(K_\nu x^\nu) = 0, \quad (7.9.38)$$

Hence, K^a is a Killing vector field with respect to η_{ab} , and $\tilde{K}^a = K^a$ is a Killing vector field with respect to g_{ab} in the linear approximation. Then, (7.9.38) gives rise to the Killing equation $\nabla_a K_b + \nabla_b K_a = 0$, where ∇_a is the torsion-free derivative operator associated with g_{ab} . Since $K_b = (\mathrm{d}\Phi)_b = \nabla_b \Phi$, the torsion-free condition of ∇_a leads to $\nabla_b K_a = \nabla_a K_b$. Thus, the Killing equation becomes

$$\nabla_a K_b = 0, \quad \text{i.e., } \nabla_a \tilde{K}^b = 0. \quad (7.9.39)$$

This indicates that the rays of these gravitational plane waves are parallel to each other. Therefore, they are called **plane-fronted gravitational waves with parallel rays**, or **pp-waves** for short.²⁰ Generally speaking, any spacetime that admits a nonzero null vector field \tilde{K}^a satisfying $\nabla_a \tilde{K}^b = 0$ is called a pp-wave, see Stephani et al. (2003).

Now we will find this wave solution explicitly. Since K^a is nonzero, in a Lorentzian coordinate system $\{x^\mu\}$ (with $t \equiv x^0$) it can be decomposed as

$$K^a = \omega(\partial/\partial t)^a + k^a, \quad (7.9.40)$$

then ω and k^a can be interpreted as the angular frequency and the wave 3-vector, respectively. Also, K^a being null indicates that $\omega^2 = k^a k_a \equiv k^2$. One can further choose $\{x^\mu\}$ such that k^a is in the z -direction ($z \equiv x^3$), i.e., the wavefront of each time t is a constant- z plane (the phase $K_\mu x^\mu = -\omega t + kz$ at t is only a function of z). Then, K^a can be expressed as

$$K^a = \omega(\partial/\partial t)^a + k(\partial/\partial z)^a, \quad (7.9.41)$$

with $\omega = \sqrt{k^2} \equiv k$. In this coordinate system, the conditions in (7.9.34) result in

$$H_{11} + H_{22} = 0, \quad H_{v3} = H_{3v} = H_{0v} = H_{v0} = 0, \quad v = 0, 1, 2, 3. \quad (7.9.42)$$

Thus, among the components $H_{\mu\nu}$ of H_{ab} , the nonvanishing ones can only be $H_{11} = -H_{22}$ and $H_{12} = H_{21}$. Therefore, H_{ab} can be written as

$$H_{ab} = H_{11} H_{ab}^{(+)} + H_{12} H_{ab}^{(\times)}, \quad (7.9.43)$$

²⁰ Notice that $g_{ab} = \eta_{ab} + f(K_\mu x^\mu) H_{ab}$ is a pp-wave only in the linear approximation.

where

$$H_{ab}^{(+)} = (dx^1)_a (dx^1)_b - (dx^2)_a (dx^2)_b, \quad (7.9.44)$$

$$H_{ab}^{(\times)} = (dx^1)_a (dx^2)_b + (dx^2)_a (dx^1)_b. \quad (7.9.45)$$

In (7.9.43), H_{11} and H_{12} are arbitrary real numbers, corresponding to the two degrees of freedom we discussed in Sect. 7.9.1 by counting the degrees of freedom. Correspondingly, if we define

$$\gamma_{ab}^{(+)} = f(K_\mu x^\mu) H_{ab}^{(+)} = f(-\omega t + kz)[(dx^1)_a (dx^1)_b - (dx^2)_a (dx^2)_b], \quad (7.9.46)$$

$$\gamma_{ab}^{(\times)} = f(K_\mu x^\mu) H_{ab}^{(\times)} = f(-\omega t + kz)[(dx^1)_a (dx^2)_b + (dx^2)_a (dx^1)_b], \quad (7.9.47)$$

then a solution in the form of (7.9.30) can be expressed as

$$\gamma_{ab} = H_{11} \gamma_{ab}^{(+)} + H_{12} \gamma_{ab}^{(\times)}. \quad (7.9.48)$$

Plugging the above solution into (7.8.5) yields the linearized Riemann curvature tensor

$$R_{acbd}^{(1)} = (K_d K_{[a} H_{c]b} - K_b K_{[a} H_{c]d}) f''(K_\mu x^\mu).$$

To verify that the curvature is indeed nonzero, we can decompose it into two terms:

$$R_{acbd}^{(1)} = H_{11} R_{acbd}^{(1)(+)} + H_{12} R_{acbd}^{(1)(\times)}, \quad (7.9.49)$$

where

$$R_{acbd}^{(1)(+)} = (K_d K_{[a} H_{c]b}^{(+)} - K_b K_{[a} H_{c]d}^{(+)}) f''(K_\mu x^\mu), \quad (7.9.50)$$

$$R_{acbd}^{(1)(\times)} = (K_d K_{[a} H_{c]b}^{(\times)} - K_b K_{[a} H_{c]d}^{(\times)}) f''(K_\mu x^\mu). \quad (7.9.51)$$

It is obvious to see that $R_{acbd}^{(1)(+)} \neq 0$ since, for example,

$$R_{acbd}^{(1)(+)} \left(\frac{\partial}{\partial x^1} \right)^d = \frac{f''(-\omega t + kz)}{2} K_b [K_c (dx^1)_a - K_a (dx^1)_c],$$

and similarly $R_{acbd}^{(1)(\times)} \neq 0$. Since $H_{ab}^{(+)}$ and $H_{ab}^{(\times)}$ are linearly independent, $R_{acbd}^{(1)} \neq 0$ if either H_{11} or H_{12} is nonzero. Therefore, for a nontrivial γ_{ab} in (7.9.48), the metric $g_{ab} = \eta_{ab} + \gamma_{ab}$ is not flat, and hence it indeed describes a gravitational plane wave. In the special case where $f(K_\mu x^\mu) = \cos(K_\mu x^\mu + \theta_0)$ with θ_0 a constant, the gravitational wave is called a **monochromatic gravitational plane wave**.

A gravitational wave of the form (7.9.46) is said to be **plus-polarized** or of **mode +**, and a gravitational wave of the form (7.9.47) is said to be **cross-polarized** or of **mode ×**. Besides the plus-polarized mode and cross-polarized mode one can also

have an arbitrary polarized mode with $H_{ab} = \alpha H_{ab}^{(+)} + \beta H_{ab}^{(\times)}$ satisfying $\alpha^2 + \beta^2 = 1$. All these polarization modes are on an equal footing. In fact, if we set

$$\begin{aligned} x'^0 &= x^0 = t, & x'^1 &= \frac{1}{\sqrt{2}}(x^1 + x^2), \\ x'^3 &= x^3 = z, & x'^2 &= \frac{1}{\sqrt{2}}(-x^1 + x^2), \end{aligned} \quad (7.9.52)$$

then $\{x'^\mu\}$ is another Lorentzian coordinate system of η_{ab} . It is easy to verify that

$$\gamma_{ab}^{(+)} = -f(-\omega t + kz) [(\mathrm{d}x'^2)_a (\mathrm{d}x'^1)_b + (\mathrm{d}x'^1)_a (\mathrm{d}x'^2)_b], \quad (7.9.53)$$

$$\gamma_{ab}^{(\times)} = f(-\omega t + kz) [(\mathrm{d}x'^1)_a (\mathrm{d}x'^1)_b - (\mathrm{d}x'^2)_a (\mathrm{d}x'^2)_b]. \quad (7.9.54)$$

Thus, in the new coordinate system $\{x'^\mu\}$, gravitational waves $\gamma_{ab}^{(+)}$ and $\gamma_{ab}^{(\times)}$ are now cross-polarized and plus-polarized, respectively, which shows that the plus-polarized and cross-polarized modes are equivalent up to a choice of the Lorentzian coordinate system.

[Optional Reading 7.9.2]

To see more precisely that all the polarization modes of a gravitational wave are on an equal footing, we define the following vector fields:

$$\begin{aligned} (e_0)^a &= (\partial/\partial t)^a, & (e_1^{(+)})^a &= (\partial/\partial x^1)^a, & (e_2^{(+)})^a &= (\partial/\partial x^2)^a, \\ (e_1^{(\times)})^a &= (\partial/\partial x'^1)^a = \frac{1}{\sqrt{2}}[(e_1^{(+)})^a + (e_2^{(+)})^a], & (e_2^{(\times)})^a &= (\partial/\partial x'^2)^a = \frac{1}{\sqrt{2}}[-(e_1^{(+)})^a + (e_2^{(+)})^a], \end{aligned}$$

where x'^μ are given in (7.9.52). Then, it is easy to verify that

$$\begin{aligned} (H^{(+)})^a_b (e_0)^b &= 0, & (H^{(+)})^a_b (e_1^{(+)})^b &= (e_1^{(+)})^a, \\ (H^{(+)})^a_b K^b &= 0, & (H^{(+)})^a_b (e_2^{(+)})^b &= -(e_2^{(+)})^a, \\ (H^{(\times)})^a_b (e_0)^b &= 0, & (H^{(\times)})^a_b (e_1^{(\times)})^b &= (e_1^{(\times)})^a, \\ (H^{(\times)})^a_b K^b &= 0, & (H^{(\times)})^a_b (e_2^{(\times)})^b &= -(e_2^{(\times)})^a. \end{aligned}$$

We can see that ① $(e_0)^a$, K^a and their linear combinations, such as $(e_3)^a = \frac{1}{\omega}K^a - (e_0)^a$, are all eigenvectors of both $(H^{(+)})^a_b$ and $(H^{(\times)})^a_b$ with eigenvalue 0; ② $(e_1^{(+)})^a$ and $(e_2^{(+)})^a$ are eigenvectors of $(H^{(+)})^a_b$ with eigenvalues ± 1 , respectively; ③ $(e_1^{(\times)})^a$ and $(e_2^{(\times)})^a$ are eigenvectors of $(H^{(\times)})^a_b$ with eigenvalues ± 1 , respectively.

For the polarization tensor H_{ab} expressed in (7.9.43), we can set $H \equiv \sqrt{(H_{11})^2 + (H_{12})^2}$ and $0 \leq \psi < \pi$ such that $H_{11} = H \cos 2\psi$ and $H_{12} = H \sin 2\psi$. Then, (7.9.43) can be written as

$$H_{ab} = HH_{ab}^{(\psi)}, \quad \text{where } H_{ab}^{(\psi)} = H_{ab}^{(+)} \cos 2\psi + H_{ab}^{(\times)} \sin 2\psi. \quad (7.9.55)$$

It is easy to see that K^a , $(e_0)^a$ and their linear combinations are all eigenvectors of $(H^{(\psi)})^a_b$ with eigenvalue 0. Moreover,

$$(e_+^{(\psi)})^a = (e_1^{(+)})^a \cos \psi + (e_2^{(+)})^a \sin \psi \quad \text{and} \quad (e_-^{(\psi)})^a = -(e_1^{(+)})^a \sin \psi + (e_2^{(+)})^a \cos \psi \quad (7.9.56)$$

are also eigenvectors of $(H^{(\psi)})^a_b$ with eigenvalues ± 1 , respectively. Especially, we have $H_{ab}^{(+)} = H_{ab}^{(0)}$ and $H_{ab}^{(\times)} = H_{ab}^{(\pi/4)}$, and correspondingly

$$(e_1^{(+)})^a = (e_+^{(0)})^a, \quad (e_2^{(+)})^a = (e_-^{(0)})^a, \quad (7.9.57)$$

$$(e_1^{(\times)})^a = (e_+^{(\pi/4)})^a, \quad (e_2^{(\times)})^a = (e_-^{(\pi/4)})^a. \quad (7.9.58)$$

Therefore, we can see clearly that $H_{ab}^{(+)}$ and $H_{ab}^{(\times)}$ are nothing but two special cases of $H_{ab}^{(\psi)}$, and all $H_{ab}^{(\psi)}$ are on an equal footing.

The geometric meaning of the $(e_\pm^{(\psi)})^a$ in (7.9.56) is clear: by rotating about the z -axis by an angle ψ , the eigenvectors $(e_1^{(+)})^a = (\partial/\partial x^1)^a$ and $(e_2^{(+)})^a = (\partial/\partial x^2)^a$ of $H_{ab}^{(+)}$ transform to $(e_+^{(\psi)})^a$ and $(e_-^{(\psi)})^a$, respectively, and become eigenvectors of $H_{ab}^{(\psi)}$. In (7.9.55), $H_{ab}^{(\psi)}$ also looks like a rotation in the “plane” containing $H_{ab}^{(+)}$ and $H_{ab}^{(\times)}$. However, if the eigenvector rotates by an angle ψ , the corresponding rotation angle of the polarization tensor is 2ψ . This indicates that the polarization tensor will come back to itself after rotating about the z -axis by (an integer times) π , which is different from the fact that the polarization of an electromagnetic wave will come back after rotating by at least 2π . This difference manifests that gravitons and photons have different spins. It is generally believed that general relativity eventually must be combined with quantum theory and become a complete and consistent quantum theory of gravity. Although until now this theory has yet to be found, physicists still often talk about the quantization of the gravitational field and its quanta—**gravitons**. Roughly speaking, the relation between gravitons and gravitational plane waves is similar to the relation between photons and electromagnetic plane waves. Gravitons have no rest mass just like photons, as they both propagate at the speed of light in vacuum, while the different rotation angles between their polarization modes is closely related to the following fact: photons have a spin of 1, while gravitons have a spin of 2.

[The End of Optional Reading 7.9.2]

Now let us discuss the physical effect of polarized gravitational waves. Consider the following monochromatic gravitational plane wave solution:

$$\gamma_{ab} = h \cos(\omega t - kz) [(dx^1)_a (dx^1)_b - (dx^2)_a (dx^2)_b]. \quad (7.9.59)$$

This is a plus-polarized gravitational wave, with the positive constants h , ω and k being the amplitude, angular frequency and wavenumber (the magnitude of the wave 3-vector) of the gravitational wave, respectively. Imagine that there are some particles in the source-free region, each labeled by a unique parameter $\varphi \in [0, 2\pi]$. Suppose the world line of the particle labeled by the parameter φ is described by the following parametric equations:

$$\begin{aligned} t &= t(\tau), & x^1 &= x^1(\tau) = a \cos \varphi, \\ z &= z(\tau) = 0, & x^2 &= x^2(\tau) = a \sin \varphi, \end{aligned} \quad (7.9.60)$$

where $a > 0$ is a constant. When there is no gravitational wave, these particles are located along a circle of radius a at rest in the reference frame of $\{x^\mu\}$. When the gravitational wave of the form (7.9.59) passes through this region, the metric becomes

$$\begin{aligned} g_{ab} = \eta_{ab} + \gamma_{ab} &= -(dt)_a(dt)_b + [1 + h \cos(\omega t - kz)](dx^1)_a(dx^1)_b \\ &\quad + [1 - h \cos(\omega t - kz)](dx^2)_a(dx^2)_b + (dz)_a(dz)_b. \end{aligned}$$

It can be proved that for the particles on the circle described by (7.9.60), their world lines are still geodesics with respect to g_{ab} . [See Exercise 7.10. In fact, the result of which shows that the t -coordinate lines for any gravitational wave of the form (7.9.30) are geodesics]. However, the coordinates x^1 and x^2 are no longer the spatial Cartesian coordinates of these particles at a time t . Instead, their spatial Cartesian coordinates at t are now

$$y^1 = x^1 \sqrt{1 + h \cos \omega t}, \quad y^2 = x^2 \sqrt{1 - h \cos \omega t}$$

and z . From the parametric equations of the world lines of these particles, we can see that they are located along an ellipse at t , described by

$$\left(\frac{y^1}{a\sqrt{1+h \cos \omega t}} \right)^2 + \left(\frac{y^2}{a\sqrt{1-h \cos \omega t}} \right)^2 = 1, \quad z = 0. \quad (7.9.61)$$

For any integer n , when $2n\pi - \frac{\pi}{2} \leq \omega t \leq 2n\pi + \frac{\pi}{2}$, the major and the minor axes of the ellipse are along the x^1 -axis and the x^2 -axis, respectively; when $2n\pi + \frac{\pi}{2} \leq \omega t \leq 2n\pi + \frac{3\pi}{2}$, the major and the minor axes of the ellipse are exchanged, along the x^2 -axis and x^1 -axis, respectively. Therefore, as the gravitational wave passes through, these particles are located along an oscillating ellipse, as shown in Fig. 7.16. The eccentricity of the ellipse at t can be calculated as

$$e_{\text{grav}}(t) = \sqrt{\frac{2h|\cos \omega t|}{1 + h|\cos \omega t|}} \cong \sqrt{2h|\cos \omega t|}. \quad (7.9.62)$$

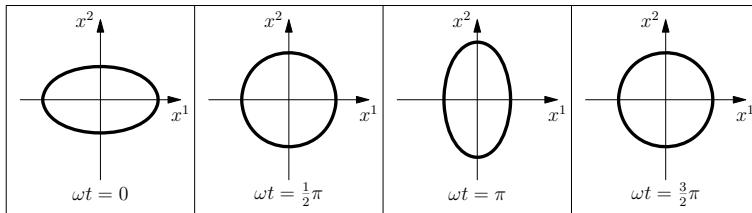


Fig. 7.16 The effect of a linearly polarized gravitational plane wave on a circle in one period

Hence, the maximum value of $e_{\text{grav}}(t)$ is $\sqrt{\frac{2h}{1+h}} \cong \sqrt{2h}$, which only depends on the amplitude h . It is important that the directions of the major and the minor axes are eigenvectors of the polarization tensor, which can be referred to as the **polarization directions** of the gravitational wave. From the viewpoint of continuum mechanics, the effect of a weak gravitational wave can be regarded as a strain.

[Optional Reading 7.9.3]

The polarization modes of gravitational waves we discussed above are analogous to the linear polarization modes of electromagnetic waves, whose polarization directions are fixed. As we know, electromagnetic waves can be circularly/elliptically polarized. Similarly, gravitational waves can also be circularly/elliptically polarized. For example, given two nonzero constants $h_{(+)}$ and $h_{(\times)}$, the gravitational wave described by

$$\gamma_{ab} = h_{(+)} H_{ab}^{(+)} \cos(-\omega t + kz) + h_{(\times)} H_{ab}^{(\times)} \sin(-\omega t + kz) \quad (7.9.63)$$

is elliptically polarized. It can be seen from (7.9.55) and (7.9.56) that when $h_{(+)}h_{(\times)} > 0$, the angular velocity of the polarization (i.e., the angular velocity of the eigenvector with eigenvalue +1) along the propagation direction is $-\omega/2$; when $h_{(+)}h_{(\times)} < 0$, the angular velocity of the polarization is $\omega/2$.

Notice that the metric $g_{ab} = \eta_{ab} + \gamma_{ab}$ with the γ_{ab} given in (7.9.63) does not abide by the ansatz (7.9.30), and thus the conclusions for (7.9.30) may not be applicable to it. However, one can show that (exercise) in the linear approximation, the spacetime corresponding to (7.9.63) is still a pp-wave, and the t -coordinate lines are still geodesics.

[The End of Optional Reading 7.9.3]

By means of the geodesic deviation equation, the effect of polarized gravitational waves can also be discussed by considering the tidal acceleration of nearby geodesic observers. In this way, one can study how a family of geodesics will be distorted by gravitational waves. This effect will be analyzed in Optional Reading 7.9.4, the discussion therein can even be applied to gravitational waves without linear approximation.

So far we have discussed wave solutions to the linearized Einstein equation. However, Einstein's equation is a nonlinear equation, and general relativity is a nonlinear theory. Although in many cases we can apply the weak-field approximation, the nonlinearity must not be ignored for a strong gravitational field. This is a significant difference between electromagnetic waves (in Minkowski spacetime) and gravitational waves. Maxwell's equations are linear equations, where the superposition principle is applicable, and so two electromagnetic waves propagating in the same space do not influence each other. In contrast, generally speaking, there exists interaction (scattering) between two gravitational waves. The collision of gravitational plane waves has been investigated in the pioneering works of R. Penrose, K. Khan and P. Szekeres, the readers may refer to d'Inverno (1992) for a review. For an example of gravitational plane waves not limited to the linear approximation, see Optional Reading 7.9.2.

[Optional Reading 7.9.4]

Now we introduce a specific example of gravitational plane waves not limited to the linear approximation [see Sachs and Wu (1977)]. Suppose $\{t, x, y, z\}$ is a Lorentzian system in

Minkowski spacetime $(\mathbb{R}^4, \eta_{ab})$. Let $u \equiv t - z$, and $f(u)$ and $g(u)$ be two arbitrary smooth functions of u with the only requirement being that $f^2 + g^2$ is nonvanishing. Suppose P is a function of the coordinates x, y and u defined as follows:

$$P(x, y, u) = \frac{1}{2}f(u)(x^2 - y^2) + g(u)xy. \quad (7.9.64)$$

It is not difficult to verify that

$$g_{ab} := \eta_{ab} + 2P(du)_a(du)_b = \eta_{ab} + 2P[(dt)_a - (dz)_a][(dt)_b - (dz)_b] \quad (7.9.65)$$

is a Lorentzian metric field on \mathbb{R}^4 . Firstly, it can be easily seen from the above equation that g_{ab} is symmetric. Secondly, let

$$K^a \equiv (\partial/\partial t)^a + (\partial/\partial z)^a, \quad (7.9.66)$$

then it is easy to verify that $g_{ab}K^aK^b = 0$, i.e., K^a is a null vector field measured by g_{ab} . Introduce a basis (tetrad) field on \mathbb{R}^4 :

$$\begin{aligned} (e_1)^a &= (\partial/\partial x)^a, & (e_2)^a &= (\partial/\partial y)^a, & (e_3)^a &= K^a, \\ (e_4)^a &= \frac{1}{2}[(\partial/\partial t)^a - (\partial/\partial z)^a] + PK^a. \end{aligned} \quad (7.9.67)$$

By a straightforward calculation (exercise) we can see that the metric components $g_{\mu\nu} \equiv g_{ab}(e_\mu)^a(e_\nu)^b$ can be arranged into the following matrix:

$$(g_{\mu\nu}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix}. \quad (7.9.68)$$

The matrix being invertible indicates that g_{ab} is non-degenerate, and thus is a metric tensor field. It is not difficult to see that it has the Lorentzian signature. The above discussion indicates that (\mathbb{R}^4, g_{ab}) is a spacetime, which has the same base manifold as \mathbb{R}^4 but has a different metric field. By calculating the curvature tensor we can see that this is a curved spacetime (see Proposition 7.9.4). The inverse matrix of (7.9.68) equipped with the basis vectors in (7.9.67) gives

$$\begin{aligned} g^{ab} &= (\partial/\partial x)^a(\partial/\partial x)^b + (\partial/\partial y)^a(\partial/\partial y)^b - (1 + 2P)(\partial/\partial t)^a(\partial/\partial t)^b \\ &\quad + (1 - 2P)(\partial/\partial z)^a(\partial/\partial z)^b - 2P[(\partial/\partial t)^a(\partial/\partial z)^b + (\partial/\partial z)^a(\partial/\partial t)^b]. \end{aligned} \quad (7.9.69)$$

We will use g^{ab} and g_{ab} to raise and lower indices.

Proposition 7.9.4 *The g_{ab} defined by (7.9.65) is a non-flat solution to the vacuum Einstein equation.*

Proof First we compute the Riemann tensor $R_{abc}{}^d$ of g_{ab} using the tetrad method introduced in Sect. 5.7. Step one: choose the tetrad in (7.9.67). It follows from (7.9.68) that this is a rigid tetrad (although not orthonormal). It is easy to verify that its dual tetrad reads

$$(e^1)_a = (dx)_a, \quad (e^2)_a = (dy)_a, \quad (e^3)_a = \frac{1}{2}[(dt)_a + (dz)_a] - P(du)_a, \quad (e^4)_a = (du)_a. \quad (7.9.70)$$

Step two: compute the connection 1-forms using Theorem 5.7.4. One finds that there are only four nonvanishing $\omega_{\mu\nu}$:

$$\begin{aligned}-\omega_{41} &= \omega_{14} = \omega_{144} e^4 = -(fx + gy)du, \\ -\omega_{42} &= \omega_{24} = \omega_{244} e^4 = -(gx - fy)du.\end{aligned}\quad (7.9.71)$$

From the inverse of (7.9.68) we can see that the components $g^{\mu\nu}$ of g^{ab} in the dual basis can also be arranged into the matrix on the right-hand side of (7.9.68), and hence it follows from $\omega_\mu^\rho = g^{\rho\nu} \omega_{\mu\nu}$ that the nonvanishing ω_μ^ρ are

$$\omega_4^1 = \omega_1^3 = (fx + gy)du, \quad \omega_4^2 = \omega_2^3 = (gx - fy)du. \quad (7.9.72)$$

The third step is to compute all the curvature 2-forms R_μ^v from ω_μ^v using Cartan's second equation of structure. Since all the nonvanishing ω_μ^ρ are shown in (7.9.72), we have $\omega_\mu^\lambda \wedge \omega_\lambda^\rho = 0$, and hence $R_\mu^v = d\omega_\mu^v$. Therefore, all the nonvanishing R_μ^v are

$$\begin{aligned}R_4^1 &= R_1^3 = f dx \wedge du + g dy \wedge du = f e^1 \wedge e^4 + g e^2 \wedge e^4, \\ R_4^2 &= R_2^3 = g dx \wedge du - f dy \wedge du = g e^1 \wedge e^4 - f e^2 \wedge e^4.\end{aligned}\quad (7.9.73)$$

Thus, we obtain the Riemann tensor

$$\begin{aligned}R_{abc}^d &= R_{ab1}^3(e^1)_c(e_3)^d + R_{ab2}^3(e^2)_c(e_3)^d + R_{ab4}^1(e^4)_c(e_1)^d + R_{ab4}^2(e^4)_c(e_2)^d \\ &= [f(e^1)_a \wedge (e^4)_b + g(e^2)_a \wedge (e^4)_b][(e^1)_c(e_3)^d + (e^4)_c(e_1)^d] \\ &\quad + [g(e^1)_a \wedge (e^4)_b - f(e^2)_a \wedge (e^4)_b][(e^2)_c(e_3)^d + (e^4)_c(e_2)^d].\end{aligned}\quad (7.9.74)$$

This is a nonvanishing tensor, since at least one of the following components is nonvanishing (the requirement for f and g is that $f^2 + g^2$ is nonvanishing):

$$R_{414}^1 = R_{abc}^d(e_4)^a(e_1)^b(e_4)^c(e^1)_d = -f, \quad R_{424}^1 = R_{abc}^d(e_4)^a(e_2)^b(e_4)^c(e^1)_d = -g.$$

This indicates that (\mathbb{R}^4, g_{ab}) is not a flat spacetime. It is easy to find the Ricci tensor from (7.9.74):

$$R_{ac} = R_{abc}^b = (f - f)(e^4)_a(e^4)_c = 0,$$

and thus g_{ab} is a solution to the vacuum²¹ Einstein equation. \square

For later use, we can also derive R_{abcd} from (7.9.74), see the following proposition:

Proposition 7.9.5

$$\begin{aligned}R_{abcd} &= [f(e^1)_a \wedge (e^4)_b + g(e^2)_a \wedge (e^4)_b](e^4)_c(e^1)_d \\ &\quad + [g(e^1)_a \wedge (e^4)_b - f(e^2)_a \wedge (e^4)_b](e^4)_c(e^2)_d.\end{aligned}\quad (7.9.75)$$

Proof Exercise 7.11. Hint: use $R_{abcd} = g_{de} R_{abc}^e$, and notice that

$$g_{de}(e_3)^e \equiv (e_3)_d = g_{3\mu}(e^\mu)_d = g_{34}(e^4)_d = -(e^4)_d, \quad g_{de}(e_1)^e \equiv (e_1)_d = g_{11}(e^1)_d = (e^1)_d.$$

\square

Given the importance of the null vector K^a in the propagation of gravitational waves, let us prove the following proposition:

²¹ In fact, this equation is $R_{ac} = -(\partial_1 \partial_1 P + \partial_2 \partial_2 P)(e^4)_a(e^4)_c$. P taking the specific form in (7.9.64) makes $\partial_1 \partial_1 P = f = -\partial_2 \partial_2 P$, which assures $R_{ac} = 0$.

Proposition 7.9.6 Suppose ∇_b is the torsion-free derivative operator associated with the g_{ab} in (7.9.65), then $\nabla_b K^a = 0$.

Proof Adopt the tetrad in (7.9.67) as well as its dual tetrad (7.9.70) and notice that $K^a = (e_3)^a$. It follows from (5.7.4) that $\omega_3{}^\nu{}_a = -\gamma^\nu{}_{3\tau}(e^\tau)_a$, $\nu = 1, 2, 3, 4$. Since the non-vanishing $\omega_\mu{}^\nu{}_a$ are shown in (7.9.72), we have $\omega_3{}^\nu{}_a = 0$, $\nu = 1, 2, 3, 4$. Thus, from the above equation we get $\gamma^\nu{}_{3\tau} = 0$, $\nu, \tau = 1, 2, 3, 4$, and hence it follows from (5.7.1) that

$$(e_\tau)^b \nabla_b (e_3)^a = \gamma^\nu{}_{3\tau} (e_\nu)^a = 0, \quad \tau = 1, 2, 3, 4.$$

Since $(e_\tau)^b$ is an arbitrary basis vector, the above equation indicates that $\nabla_b (e_3)^a = 0$. Noticing that $(e_3)^a = K^a$, we have $\nabla_b K^a = 0$. \square

From Proposition 7.9.6 we obtain $K^b \nabla_b K^a = 0$ and $\nabla_{(a} K_{b)} = 0$, and thus ① the integral curves of K^a are (null) geodesics; ② K^a is a Killing vector field.

The above discussions are purely mathematical. Physically speaking, the curved spacetime (\mathbb{R}^4, g_{ab}) represents a gravitational plane wave. It follows from (7.9.65) that P is the only available quantity that determines (\mathbb{R}^4, g_{ab}) , and thus the first thing we should investigate when studying gravitational waves is the function $P(x, y, u)$. To facilitate understanding, we first look at a simple example. Suppose $f(u)$ and $g(u)$ can be expressed as

$$f(u) = F \cos \omega u, \quad g(u) = G \cos \omega u, \quad (7.9.76)$$

where F, G and ω are positive constants, then

$$2P(x, y, u) = [F(x^2 - y^2) + Gxy] \cos(\omega t - kz) \quad (\text{where } k \equiv \omega). \quad (7.9.77)$$

The allure of the above equation is that it looks like some kind of monochromatic plane wave. However, notice that although $(\partial/\partial t)^a$ and $(\partial/\partial z)^a$ are respectively timelike and spacelike vector field when measured by η_{ab} , this is not necessarily true when measured by g_{ab} . If $(\partial/\partial t)^a$ were not timelike or $(\partial/\partial z)^a$ were not spacelike, one could not treat t and z as time and spatial coordinates, and the wave interpretation of (7.9.77) would become unclear. Fortunately, it can be proved that there indeed exist certain spacetime regions in (\mathbb{R}^4, g_{ab}) , where $(\partial/\partial t)^a$ and $(\partial/\partial z)^a$ are timelike and spacelike when measured by g_{ab} , and thus at least in these regions we can interpret (7.9.77) as a monochromatic gravitational plane wave propagating along the z -direction at the speed of light $c = 1$. The product of K^a defined by (7.9.66) and ω can be interpreted as the wave 4-vector ωK^a , since (7.9.66) indicates that the time and spatial components of ωK^a in the coordinate system $\{t, x, y, z\}$ are the angular frequency ω and the wave 3-vector \vec{k} measured in this system:

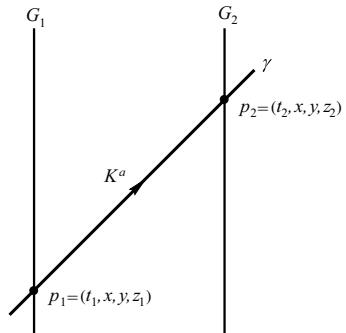
$$\omega K^0 = \omega, \quad \omega K^1 = \omega K^2 = 0, \quad \omega K^3 = k = \omega.$$

K^a (and hence ωK^a) being null reflects the fact that the phase ωu of the above gravitational wave propagates at the speed of light, see Fig. 7.15 (in which K^a should be substituted by ωK^a). Suppose G_1 and G_2 are two inertial observers (measured by η_{ab}), whose spatial coordinates are (x, y, z_1) and (x, y, z_2) , respectively. They have different phases at the time t_1 , which are $\omega t_1 - kz_1$ and $\omega t_1 - kz_2$. Suppose after some amount of time $t_2 - t_1$, G_2 “acquires” the phase of G_1 at t_1 , i.e.,

$$\omega t_2 - kz_2 = \omega t_1 - kz_1,$$

then we say that the value of the phase $\omega t_1 - kz_1$ propagates from G_1 to G_2 in a time interval $t_2 - t_1$, and so the speed of the propagation is

Fig. 7.17 The phase value $\omega t_1 - kz_1$ of the observer G_1 at t_1 propagates to G_2 after $t_2 - t_1$



$$v = \frac{z_2 - z_1}{t_2 - t_1} = \frac{\omega}{k} = 1.$$

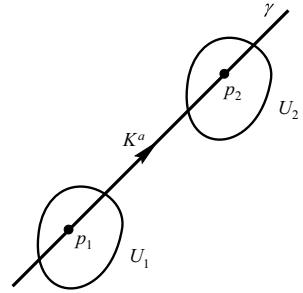
Thus, the speed of the propagation of gravitational waves is the speed of light. (This is only the coordinate speed, what is more meaningful in the geometric language is the phase velocity. The waveform being null in the 4-dimensional language assures that this phase velocity is the speed of light). Figure 7.17 is a 4-dimensional illustration of this discussion, in which γ is an integral curve of the null 4-vector K^a (a null geodesic), and p_1 and p_2 are the intersections of γ and the world lines of G_1 and G_2 . The phase value $\omega t_1 - kz_1$ at p_1 is “acquired” by G_2 at p_2 : the phase propagates from p_1 to p_2 along the null geodesic. Note that the physical interpretation above can only apply to some certain regions of (\mathbb{R}^4, g_{ab}) , where $(\partial/\partial t)^a$ is timelike and $(\partial/\partial z)^a$ is spacelike. However, now we can pull out the non-intrinsic factors such as observers and coordinates and only leave the null geodesic γ and two arbitrary points p_1 and p_2 on it. In this way, the wave interpretation can be carried over to the whole spacetime. In fact, K^a represents the direction of the propagation of all the information (not only the phase) of the gravitational wave. The reason is as follows: as K^a is a Killing vector field, its corresponding one-parameter group of diffeomorphisms is a one-parameter group of isometries, and the integral curves of K^a are exactly the orbits of this isometry group. Suppose U_2 is an arbitrary neighborhood of p_2 (see Fig. 7.18), then there must exist a neighborhood U_1 of p_1 and an isometry $\phi : U_1 \rightarrow U_2$ such that $p_2 = \phi(p_1)$. Therefore, any information about the gravitational wave in U_2 is completely contained in U_1 (due to the isometry). In this sense, we can say that all the information of the gravitational wave propagates along K^a (at the speed of light). This interpretation based on the isometries can be applied to not only the special case in (7.9.76), but also the g_{ab} defined by (7.9.64) [in which $f(u)$ and $g(u)$ are arbitrary] and (7.9.65). Hence, we say that there exists a gravitational plane wave in the spacetime (\mathbb{R}^4, g_{ab}) , or refer to (\mathbb{R}^4, g_{ab}) as a **gravitational plane wave spacetime**. Sachs and Wu (1977) also provides a deeper argument for this gravitational plane wave interpretation from the perspective of group theory by comparing it with the electromagnetic plane waves in Minkowski spacetime. Furthermore, Proposition 7.9.6 indicates that g_{ab} is a pp-wave.²² When f and g are linearly dependent, then (\mathbb{R}^4, g_{ab}) is called a **monochromatic gravitational plane wave spacetime**.

²² In fact, the metric for any pp-wave can be expressed in the Brinkmann coordinate system in the following general form:

$$ds^2 = 2P(u, x, y)du^2 - 2dudv + dx^2 + dy^2,$$

where P is an arbitrary smooth function. It is not difficult to see that this is equivalent to (7.9.65) (by setting $v = \frac{t-z}{2}$), and taking P to be of the form (7.9.64) is just a special case.

Fig. 7.18 K^a carries the information of the gravitational wave in U_1 to U_2 faithfully



To further understand the gravitational wave of (\mathbb{R}^4, g_{ab}) , we supplement the above with the following propositions and remarks. For generality, we do not put any constraint on the form of the function $P(x, y, u)$ in the following two propositions.

Proposition 7.9.7 *Let ∇_a represent the derivative operator associated with g_{ab} in (7.9.65), then*

$$\nabla^a \nabla_a P = \frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2}. \quad (7.9.78)$$

Proof Exercise 7.14. □

Remark 1 Given a function $Q(t, x, y, z)$ in Minkowski spacetime, then

$$\partial^\mu \partial_\mu Q(t, x, y, z) = Q(t, x, y, z) \quad (7.9.79)$$

in mathematical physics is called a wave equation (with source) for the function $P(t, x, y, z)$, the physical quantity $P(t, x, y, z)$ satisfying this equation represents some kind of wave motion. $\nabla^a \nabla_a P$ on the left-hand side of (7.9.78) can also be written as $g^{\mu\nu} \nabla_\mu \nabla_\nu P$, when $g_{ab} = \eta_{ab}$ it goes back to $\partial^\mu \partial_\mu P$. Thus, $\nabla^a \nabla_a P$ is the generalization of $\partial^\mu \partial_\mu P$ in curved spacetime, and hence (7.9.78) represents some kind of wave motion of the physical quantity $P(x, y, u)$ in the curved spacetime (\mathbb{R}^4, g_{ab}) . When P takes the form of (7.9.64), we have

$$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2} = 0,$$

and hence $\nabla^a \nabla_a P = 0$, i.e., the $P(x, y, u)$ in (7.9.64) is a solution to the source-free wave equation in curved spacetime. Together with $R_{ac} = 0$ (i.e., g_{ab} satisfies the vacuum Einstein equation), we can see the legitimacy of the statement “the curved spacetime (\mathbb{R}^4, g_{ab}) represents a gravitational wave in vacuum”. This also shows (at least partially) the motivation for taking P to be of the form in (7.9.64).

Proposition 7.9.8 *The constant- u surfaces in (\mathbb{R}^4, g_{ab}) are null hypersurfaces.*

Proof It follows from (7.9.67) that $K_a = g_{ab} K^b = g_{ab} (e_3)^b$. Following the derivation in (2.6.10a) we get $g_{ab} (e_3)^b = g_{3\mu} (e^\mu)_a$, and hence

$$K_a = g_{3\mu} (e^\mu)_a = g_{34} (e^4)_a = -(e^4)_a = -\nabla_a u,$$

where in the last step we used (7.9.70). Noticing that $\nabla_a u$ is a normal covector of a constant- u surface, we can see that its normal vector $\nabla^a u = -K^a$ is null. □

Remark 2 In the special case of (7.9.76), $\omega u = \omega t - kz$ represents the phase of the wave, while ω is a constant, and hence a constant- u surface is a 3-dimensional wavefront \mathcal{S} in the 4-dimensional language. \mathcal{S} being a null hypersurface indicates that the gravitational wave in (7.9.76) propagates at the speed of light. Proposition 7.9.8 guarantees that the constant- u surfaces are still null hypersurfaces (still have K^a as the normal vector) for general $P = P(x, y, u)$. Therefore, one may regard u as some kind of (generalized) phase, and the constant- u surfaces being hypersurfaces indicates that the phase velocity of the gravitational wave represented by this general $P(x, y, u)$ is still the speed of light.

[The End of Optional Reading 7.9.4]

7.9.3 Emission of Gravitational Waves

Now we introduce the emission of gravitational waves. First, let us make a comparison with electromagnetic waves. If a charged particle in a system undergoes a non-uniform velocity (relative to an inertial frame), it will emit electromagnetic waves. As is well-known, the major contribution to the radiation field comes from electric dipole radiation, which is much stronger than the magnetic dipole radiation and electric quadrupole radiation (these two are of the same order). Similarly, under the Newtonian approximation, if a point mass in a system undergoes a non-uniform velocity, it will emit gravitational waves. What corresponds to the electric dipole moment is the **mass dipole moment**

$$\vec{D} = \sum_P m_p \vec{r}_P , \quad (7.9.80)$$

where m_P and \vec{r}_P are the mass and position vector of the point mass P , and the right-hand side of the above equation is summed over all the point masses in the system. Since the intensity of electric dipole radiation is proportional to the square of the second order time derivative of the electric dipole moment, one may expect that the contribution from the mass dipole moment to the intensity of gravitational radiation is proportional to $\ddot{\vec{D}}$. However, from (7.9.80) we can see that $\dot{\vec{D}} = \sum_P m_p \dot{\vec{r}}_P$ is equal to the total momentum \vec{p} of the system; it follows from the conservation of momentum that $\dot{\vec{p}} = 0$, and thus $\ddot{\vec{D}} = 0$, i.e., gravitational waves do not include gravitational dipole radiation corresponding to electric dipole radiation. According to the theory of electromagnetic radiation, the intensity of magnetic dipole radiation is proportional to the square of the second order time derivative of the magnetic dipole moment. The quantity in a gravitational system corresponding to the magnetic dipole moment is

$$\vec{\mu} = \sum_P \vec{r}_P \times (m_P \vec{u}_P) ,$$

where \vec{u}_P is the velocity of the point mass P , and $m_P \vec{u}_P$ is the current contribution of P . The right-hand side of the above equation is nothing but the total angular

momentum of the system. It follows from the conservation law of angular momentum that $\dot{\vec{\mu}} = 0$, and hence gravitational waves do not include gravitational dipole radiation corresponding to magnetic dipole radiation either. In short, there does not exist any dipole radiation in gravitational waves. One can only get a nonvanishing result when studying quadrupole radiation [see Misner et al. (1973) pp. 974–978 for details]. Since the order of quadrupole radiation is higher than dipole radiation, the gravitational waves emitted from a gravitational system are weaker than the electromagnetic waves emitted by an electromagnetic system in a similar condition.

The source emitting a strong gravitational wave is usually considered to be related to a dramatic change of an astrophysical or cosmological process, such as the collapse of a star that is not spherically symmetric,²³ a supernova explosion (see Sect. 9.3.2), the dramatic disturbance inside an active galactic nucleus, the merger of a pair of black holes or neutron stars, cosmic inflation (see Chap. 15), etc. [See Cai et al. (2017) for a review of different sources of gravitational waves]. In these cases the gravitational field is not weak, and thus the linear approximation is not applicable. The rigorous analysis of these processes must involve the arduous task of solving the nonlinear Einstein equation in a non-spherically symmetric case. The emission of gravitational waves is still a problem that has not been fully comprehended. Nowadays, the understanding of this problem has been furthered with the help of numerical analysis and computational simulation, which has developed into an important branch called numerical relativity.

7.9.4 *Detection of Gravitational Waves*

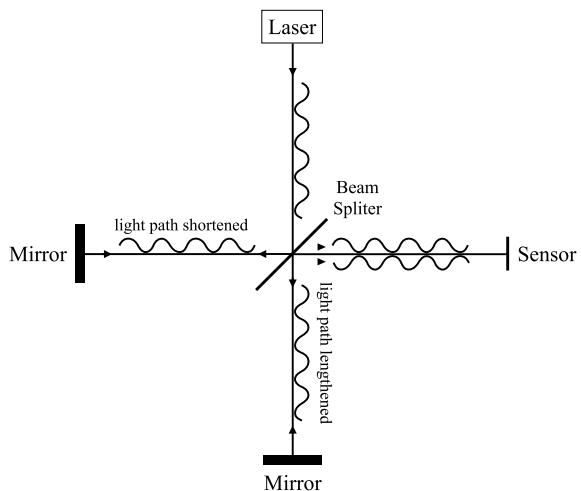
Since general relativity predicts the physical existence of gravitational radiation, the detection of gravitational waves becomes a significant subject. As we have discussed, sources for the gravitational waves that reach the solar system are all very far away. Hence, the gravitational waves being detected can be totally regarded as plane waves, and they are so weak that the linear approximation is applicable. Unfortunately, this also makes it very difficult to directly detect a gravitational wave on or near the Earth. (The currently observed gravitational waves have amplitudes as small as $h \sim 10^{-21}$). Due to such a difficulty, there were no direct observations of gravitational waves until 2015, although Joseph Weber initiated the detection of gravitational waves early in the 1960s. In the 20th century, evidence of the existence for gravitational waves merely came from indirect detections, among which the most important one is the observation of binary pulsars.

²³ According to Birkhoff's theorem (see Sect. 8.3.3), the spherical evolution of any spherically symmetric star (such as collapse and oscillation) will not emit a gravitational wave no matter how dramatic it is, just like there does not exist a spherically symmetric electromagnetic wave in Maxwell's theory. (The spherical wave of an oscillating electric dipole in a distant region is not a spherically symmetric electromagnetic wave, since the fields \vec{E} and \vec{B} are not spherically symmetric. In fact, a spherically symmetric electromagnetic wave corresponds to the radiation of an electric monopole, but this kind of radiation does not exist in Maxwell's theory).

A pulsar is a rapidly rotating neutron star (see Sect. 9.3.2), which has a mechanism of emitting electromagnetic waves. If the Earth lies in the sweeping range of a beam of radiation, then one can receive radio pulse signals with a precise period. An approximately isolated gravitational system formed by two stars orbiting around their center of mass is called a **binary star**, which emits gravitational waves due to the accelerating motion of the two stars. Like electromagnetic waves, gravitational waves carry energy and momentum as well as angular momentum when they are emitted. As a consequence, the radii of the orbits of the stars become smaller and smaller, and the period becomes shorter and shorter. However, unlike many other astrophysical processes, the emission of gravitational waves from a binary system is very weak, and so the linearized theory of gravity can be applied to calculate the loss of energy and the change of the orbital period. In order to be detectable, these effects need to satisfy at least two conditions: ① the orbit is sufficiently small (i.e., the two stars are close enough), such that the effect of general relativity is evident; ② A method for measuring the orbital period with rather high precision is available. The binary pulsar PSR 1913+16 discovered by R. A. Hulse and J. H. Taylor in 1974 happens to satisfy these two conditions. [A binary pulsar is a binary that contains a pulsar, PSR is the identifier for pulsars, while 1913 and +16 stands for its right ascension and declination (angular coordinates)]. The maximum distance between the two stars in this binary is only about 3×10^9 m (about 4.8 solar radii) which satisfies the condition ①; the pulsar in the binary makes it satisfy the condition ②: since the period of the radio signal emitted from a pulsar is reputed to be “as precise as the tick of a clock”, one can use this to record how its orbital period changes, and compare with the result calculated from general relativity. If the observation agrees with the calculation on account of gravitational waves, it will be evidence for the existence of gravitational waves. Taylor and collaborators carried out this observation with extraordinarily high accuracy and obtained the rate of change of the orbital period. After thousands of observations, their results were announced in 1978, which agrees very well with the predictions calculated from the quadrupole radiation formula in the linearized theory of gravity. This was the first quantitative evidence of gravitational waves ever since gravitational waves were proposed, even though it was indirect evidence. Hulse and Taylor were awarded the 1993 Nobel Prize in Physics for this discovery.

The first attempt to directly detect gravitational waves was started by Joseph Weber at the University of Maryland in 1966. He designed a **resonant mass antenna** for detecting gravitational waves, called the **Weber bar**. It is a suspended aluminum cylinder with length 153 cm and diameter 66 cm, which has a resonance frequency of 1660 Hz. When a gravitational wave near the resonant frequency passes through the Weber bar in a proper direction, the resonance of the bar will be excited, which will amplify the vibration and could potentially be detected by piezoelectric sensors if the change of the bar’s length is large enough. After years of efforts, Weber announced that the evidence of gravitational waves was observed from the detectors in two different locations. Unfortunately, Weber’s observation could not be confirmed by the experiments of any other group [Ohanian and Ruffini (1994); Liu and Zhao (2004)].

Fig. 7.19 Schematic diagram of an interferometric detector. The light paths along the two arms change slightly when a gravitational wave comes by, which causes the interference of the composite signal



In the 1970s, there appeared another important type of gravitational wave detector, namely a **laser interferometer**. An interferometer has two long orthogonal arms, and the idea is similar to the resonant mass antenna, i.e., to detect the length perturbations of its arms due to gravitational waves. However, the range of the detectable frequencies of a laser interferometer is much wider instead of only near a resonance frequency. Here we briefly review the principle of interferometers. An interferometer consists of two mirrors and a beam splitter (see Fig. 7.19). When a laser beam is shot to the beam splitter through the vertical arm in Fig. 7.19, part of it will be transmitted while the remaining part will be reflected, and thus the laser will be divided into two beams which propagate along the two arms of the interferometer. Each of the two beams hits the mirror placed at the end of each arm and gets bounced back to the beam splitter. After that the beams recombine and propagate towards the right through the horizontal arm in Fig. 7.19, which will be received by the sensor at the end of the horizontal arm. When there is no gravitational wave, the recombined beams are tuned to have opposite phases (a crest meets a trough) by applying a waveplate, so that the composite signal vanishes. However, when a gravitational wave comes by, the lengths of the arms will change slightly (similar to the effect shown in Fig. 7.16) and so the light paths of the two beams will change slightly, causing a nonvanishing composite signal to be received by the sensor, called an interference signal. Therefore, interference will be present when there is a gravitational wave passing by.

Based on the above idea, the study groups in MIT and Caltech started to jointly build the **Laser Interferometer Gravitational-Wave Observatory (LIGO)** since the 1980s (early discussions and attempts on interferometric detectors began in the late 1960s). After decades of preparation, LIGO started its first operation in 2002. However, it was not sensitive enough to detect any gravitational wave successfully. In 2010, LIGO was shut down and upgraded into an improved version—Advanced LIGO, whose sensitivity is about ten times its previous version. The operation of

LIGO restarted in 2015. At 09:50:45 UTC on 14 September 2015, LIGO made the first direct observation of gravitational waves [Abbott et al. (2016)]. The signal of this event was named GW150914, which comes from a merger of two black holes occurred 1.3 billion light-years away, with the amplitude of $\gamma_{\mu\nu}$ (the components of γ_{ab} in a Lorentzian coordinate system) being so small that it is equivalent to changing a length of 4 km by a thousandth of the width of a proton. Due to this unprecedented observation, three leaders of LIGO, Rainer Weiss, Barry Barish and Kip Thorne, were awarded the 2017 Nobel Prize in Physics.

To make precise detections, the LIGO observatory consists of two identical interferometers, located in Washington state and Louisiana state, USA, respectively. The distance between them is about 3030 km over the Earth's surface (the straight line distance is about 3002 km). Besides making independent measurements, an important utility of having two detectors far apart is to determine the location of the source of the gravitational wave. Since the gravitational wave travels at the speed of light, it would take 10 ms to propagate from one LIGO interferometer to the other. In the GW150914 event, the time delay between the two detectors was 7 ms. Using this time delay, the source of the signal can be located through triangulation. This is exactly the principle of how human ears identify the location of the source of a sound wave. Interestingly, the signal of GW150914 has a frequency varying between 35 Hz–250 Hz, which happens to be inside the human audible range. In 2017, the Virgo interferometer in Italy started to detect gravitational waves which provides “a third ear” for locating the source of the gravitational wave more precisely. Furthermore, having two identical LIGO detectors also helps to extract the actual gravitational wave signal from the noise. Since the detectors are extremely sensitive, any vibration from the local environment will be recorded, and one of the challenges of the detection is to remove these noises. By comparing the signals obtained by the two detectors located far apart, one can filter out the random vibrations that do not happen at both places, with the gravitational wave signals that are identical remaining. To minimize the noises, LIGO also applied a series of mechanisms to isolate the vibrations, including optics suspensions and seismic isolation, and many techniques in the data analysis, such as matched filtering. The reader may refer to Saulson (2017) for more technical details of noise reduction.

Since the first direct observation in 2015, there have already been numerous events of direct observation of gravitational waves, mainly detected by LIGO and Virgo. Nevertheless, now there are more and more gravitational wave detectors becoming available or under preparation. For example, KAGRA (Kamioka Gravitational Wave Detector) started its observation in 2020. Also, third-generation interferometric detectors with longer arms and a greater sensitivity, such as the Einstein Telescope and Cosmic Explorer, have been proposed and are expected to be available in the 2030s. Besides the ground-based interferometers, there are also multiple on-going projects for space-based interferometric detectors, such as LISA (Laser Interferometer Space Antenna), TianQin, Taiji, and DECIGO (Deci-hertz Interferometer Gravitational wave Observatory), where the long arms are replaced by the laser beams between spacecrafts. Once available, they will be used to detect low-frequency gravitational waves. In addition to interferometric detectors, there are also other methods

Table 7.1 Methods of gravitational wave detection and their frequency bands

Frequency band	Frequency range	Detection method	Current and future observatories
High-frequency	$10 \text{ Hz} \text{--} 10^6 \text{ Hz}$	Ground-based interferometer	LIGO, Virgo, KAGRA, Einstein Telescope, Cosmic Explorer
Low-frequency	$10^{-7} \text{ Hz} \text{--} 10 \text{ Hz}$	Space-based interferometer	LISA, TianQin, Taiji, DECIGO
Very-low-frequency	$10^{-10} \text{ Hz} \text{--} 10^{-7} \text{ Hz}$	Pulsar timing array	IPTA
Extremely-low-frequency	$10^{-18} \text{ Hz} \text{--} 10^{-14} \text{ Hz}$	CMB polarization	BICEP, AliCPT

of detecting gravitational waves, such as by using pulsar timing arrays [e.g., IPTA (International Pulsar Timing Array)] one can detect very-low-frequency gravitational waves, and by measuring the polarization pattern of the cosmic microwave background (CMB) [e.g., BICEP (Background Imaging of Cosmic Extragalactic Polarization), AliCPT (Ali CMB Polarization Telescope)] one can detect extremely-low-frequency gravitational waves, including the primordial gravitational waves generated in the early universe (see Sect. 10.3). For a detailed introduction to the methods of the pulsar timing array and CMB polarization, see, for example, Maggiore (2018). The above-mentioned detecting methods and their corresponding frequency bands are summarized in Table 7.1 [see also Chen et al. (2017)].

The observation of gravitational waves is significant not only because it confirmed the last undetected prediction of general relativity, but more importantly, it also opened up a brand new window for observing the universe. Traditionally, people could only make astronomical observations by detecting the electromagnetic waves in different frequency bands. Now that gravitational waves can also be directly detected, it enables more possibilities for astronomical observation. For example, since the electromagnetic field interacts with matter, the electromagnetic waves from a distant celestial object can be easily scattered or absorbed during the propagation. However, the interaction between gravitational waves and matter is much more weaker, so it is possible to observe celestial events we could not observe before (like the binary black hole merger of GW150914). Furthermore, the earliest electromagnetic radiation we can observe is the cosmic microwave background radiation when photon decoupling occurred (see Sect. 10.3), but through gravitational waves it is now possible to make observations of the early universe. With these prospects, **gravitational-wave astronomy** is currently emerging, and hopefully it will lead to more revolutionary discoveries of the universe in the near future.

[Optional Reading 7.9.5]

Using the example of gravitational plane waves in Optional Reading 7.9.2, we will now introduce the mechanism of receiving gravitational waves in a geodesic reference frame (where the world lines of the observers are geodesics) [see also Sachs and Wu (1977)]. In a

vibrating mechanical detector like a Weber bar, each molecule of the aluminum bar can be considered as an observer, and the bar can be viewed as a reference frame in a sub-spacetime of (\mathbb{R}^4, g_{ab}) . Since there also exists non-gravitational interactions between the molecules, the world lines of the molecules are not geodesics. However, in practice one can still use a geodesic reference frame (which is the simplest choice). This is because the response to the gravitational waves in the reference frame of the bar can be derived from the response in the geodesic frame through Newtonian mechanics and solid state physics [see Weber (1961)].

The relative acceleration of two neighboring observers in a geodesic reference frame under the action of the spacetime curvature is the tidal acceleration (see Sect. 7.6). Under the action of the gravitational wave in (7.9.77), the magnitude and direction of the tidal acceleration will change periodically, which leads to a relative oscillation between two neighboring observers. Take a geodesic $\gamma(\tau)$ as the fiducial observer, let us compute the tidal 3-acceleration a^c of the neighboring observers around this observer. Suppose $p \in \gamma$, Z^a is the 4-velocity of γ at p (namely the unit tangent vector of γ), and W_p is the 3-dimensional subspace in the tangent space V_p of p which is orthogonal to Z^a (in a picture it would be a small plane orthogonal to Z^a), then a spatial separation vector w^a represents a neighboring observer (Sect. 7.6).²⁴ The tidal acceleration a^c of the observer corresponding to w^a relative to the fiducial observer $\gamma(\tau)$ is given by the geodesic deviation equation (7.6.8):

$$a^c = -R_{abd}{}^c Z^a w^b Z^d. \quad (7.9.81)$$

$\forall w^b \in W_b$, the above equation determines an $a^c \in W_p$, and thus the above equation defines a linear map $\psi : W_p \rightarrow W_p$. From the “multifaceted view of tensors” (see Sect. 2.4) we can see that ψ can be viewed as a tensor of type $(1, 1)$ on W_p , denoted by $\psi^c{}_b$, i.e.,

$$a^c = \psi^c{}_b w^b. \quad (7.9.82)$$

Comparing with (7.9.81) yields

$$\psi^c{}_b = -R_{abd}{}^c Z^a Z^d. \quad (7.9.83)$$

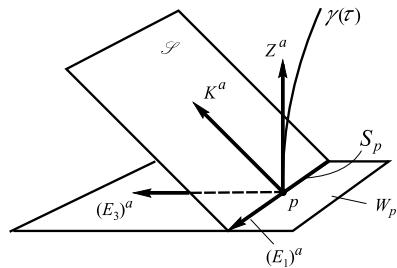
In order to compute $\psi^c{}_b$, one can first choose a convenient orthonormal triad $\{(E_i)^a\}$:

$$\begin{aligned} (E_1)^a &= (\partial/\partial x)^a + E^{-1} Z_1 K^a, \\ (E_2)^a &= (\partial/\partial y)^a + E^{-1} Z_2 K^a, \\ (E_3)^a &= E^{-1} K^a - Z^a, \end{aligned} \quad (7.9.84)$$

where $E \equiv -g_{ab} Z^a K^b > 0$, $Z_1 \equiv g_{ab} Z^a (\partial/\partial x)^b = Z_b (\partial/\partial x)^b$ (and hence Z_1 is a coordinate component of Z_b instead of a frame component), and $Z_2 \equiv g_{ab} Z^a (\partial/\partial y)^b = Z_b (\partial/\partial y)^b$. The reader should verify that: ① $\{(E_i)^a\}$ is indeed orthogonal measured by g_{ab} ; ② $(E_3)^a$ is the result of normalizing $h^a{}_b K^b = K^a + Z^a Z_b K^b$, namely the projection of K^a at p onto W_p ; ③ $\{(E_i)^a\}$ is parallelly transported (and thus is Fermi transported) along a geodesic. [Hint for the proof: It follows from $\gamma(\tau)$ being geodesic and $\nabla_a K^b = 0$ that E is a constant along the curve, from which one can easily show that $Z^b \nabla_b (E_3)^a = 0$. Noticing that $\nabla_b (\partial/\partial x)^a = -K^a \omega_1{}^3{}_b$, one can show that $Z^b \nabla_b (E_1)^a = 0$.] Let S be the wavefront that includes $p \in \gamma$ (see the null hypersurface in Fig. 7.20), $\hat{\mathcal{S}}$ represent the 3-dimensional subspace formed by all the elements in V_p tangent to \mathcal{S} , and $S_p \equiv \hat{\mathcal{S}} \cap W_p = \{w^a \in W_p \mid g_{ab} w^a K^b = 0\}$, then $\{(E_1)^a, (E_2)^a\}$ is a basis of S_p . Since in a picture we always draw a subspace (e.g., W_p) as a small plane (draw a subspace of V_p as a subspace of M), there is no difference between $\hat{\mathcal{S}}$ and \mathcal{S} in Fig. 7.20. The physical meaning

²⁴ More precisely, w^a only gives the direction of the “separation”, it is really $w^a \Delta s$ (where Δs is small) that determines a neighboring observer in this direction, see Sect. 7.6.

Fig. 7.20 In the view of a geodesic observer $\gamma(\tau)$, the gravitational wave passes by along the spatial direction $(E_3)^a$ at p , the wavefront S_p is orthogonal to $(E_3)^a$



of the mathematical settings above is very clear: in the view of a geodesic observer $\gamma(\tau)$, the gravitational wave passes by along the spatial direction $(E_3)^a$, and the 2-dimensional wavefront S_p is orthogonal to the direction of propagation $(E_3)^a$ (see Fig. 7.20). The components of ψ^c_b in the triad $\{(E_i)^a\}$ are

$$\psi^i_j = \psi^c_b (E^i)_c (E_j)^b = \psi_{cb} (E^i)^c (E_j)^b = \psi_{cb} (E_i)^c (E_j)^b = -R_{abcd} Z^a (E_j)^b Z^c (E_i)^d, \quad (7.9.85)$$

where we used the property $(E^i)^c = \delta^{ij} (E_j)^c = (E_i)^c$ of an orthonormal frame. Plugging (7.9.84) and the R_{abcd} in (7.9.75) into the equation above yields the matrix of ψ^i_j :

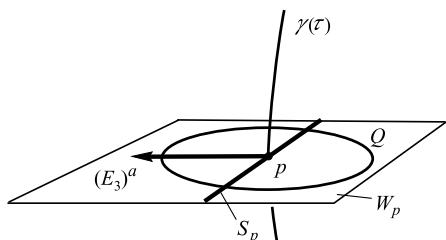
$$(\psi^i_j) = \begin{bmatrix} \alpha & \beta & 0 \\ \beta & -\alpha & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \alpha \equiv -E^2 f, \quad \beta \equiv -E^2 g. \quad (7.9.86)$$

The derivation of the above equation is left as Exercise 7.13. Hints:

- (1) Make use of $(e^1)_a (E_2)^a = (e^2)_a (E_1)^a = (e^4)_a (E_1)^a = (e^4)_a (E_2)^a = 0$;
- (2) $(e^4)_a Z^a = g_{ab} Z^a (e^4)^b = g_{ab} Z^a g^{43} (e_3)^b = -g_{ab} Z^a K^b = E$, where g^{43} is a component of g^{ab} in the frame $\{(e^\mu)_a\}$, see (7.9.68);
- (3) $(e^4)_a (E_3)^a = (e^4)_a [E^{-1} (e_3)^a - Z^a] = -(e^4)_a Z^a = -E$.

Now we discuss the physical meaning of (7.9.86). Suppose $\gamma(\tau)$ is the fiducial observer, $p \in \gamma$, and Q is the sphere orthogonal to the “small plane” W_p with a small radius whose center is p , then each point on the sphere can be viewed as the behavior of a neighboring observer at the moment p (see Fig. 7.21). Using (7.9.82) and (7.9.86), let us discuss the tidal acceleration a^c of these neighboring observers relative to $\gamma(\tau)$ under the action of gravitational waves. Each point on the sphere corresponds to a w^b . Suppose its components in the orthonormal triad $\{(E_i)^a\}$ are w^1, w^2, w^3 , then the column matrix constituted by the components of its 3-acceleration is

Fig. 7.21 Each point on the small sphere Q in the orthogonal plane W_p with p as the center represents the behavior of a neighboring observer at the same time as the event p



$$\begin{bmatrix} a^1 \\ a^2 \\ a^3 \end{bmatrix} = \begin{bmatrix} \alpha & \beta & 0 \\ \beta & -\alpha & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w^1 \\ w^2 \\ w^3 \end{bmatrix}. \quad (7.9.87)$$

If $w_1 = w_2 = 0$, i.e., w^a is parallel to $(E_3)^a$, the direction of the gravitational wave propagation, then it follows from (7.9.87) that $a^1 = a^2 = a^3 = a^4 = 0$, and thus this kind of neighboring observer has no 3-acceleration at all. This is a physical manifestation of the transverseness of gravitational waves: a neighboring observer in the longitudinal direction (which is parallel to the direction of propagation) will experience nothing, and only the neighboring observers in the transverse direction will be affected. In other words, all the tidal accelerations are orthogonal to the direction of propagation $(E_3)^a$, and thus lie in the wavefront S_p in Fig. 7.20. Hence, we only care about the transverse response, i.e., simplify (7.9.87) as

$$\begin{bmatrix} a^1 \\ a^2 \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \beta & -\alpha \end{bmatrix} = \begin{bmatrix} w^1 \\ w^2 \end{bmatrix}, \quad (7.9.88)$$

i.e., only care about the response of the points on a small circle in the 2-dimensional subspace spanned by $(E_1)^a$ and $(E_2)^a$. Take 8 representative points A, B, C, D, E, F, G, H on the circle (see Fig. 7.22). Let us discuss the following two special cases: (a) $\beta \equiv 0, \alpha > 0$; (b) $\alpha \equiv 0, \beta > 0$. From a straightforward calculation one can obtain the results in Table 7.2 and Fig. 7.22. The deformation shown in Fig. 7.22 is called a shear (see Sect. 14.1 for details).

Table 7.2 and Fig. 7.22 only reflect the tidal acceleration of the circle (and the trend of its deformation) at a certain moment. To figure out the situation of the deformation (oscillation) of the circle in a period of time, one needs the specific form of the functions $f(u)$ and $g(u)$. We still only discuss the case of $f(u) = F \cos(\omega t - kz)$ and $g(u) = G \cos(\omega t - kz)$. Equation (7.9.86) indicates that the direct factors that determine the tidal acceleration are $E^2 f$ and $E^2 g$ instead of f and g . However, it follows from K^a being geodesic (see below Proposition 7.9.6) that E is a constant on the geodesic $\gamma(\tau)$, and thus what the tidal acceleration reflects is also the values of f and g . Moreover, since the u on the geodesic $\gamma(\tau)$ and the proper time τ have a linear relation $du/d\tau = E$ (the proof is left as an exercise), the a^i - τ curve measured by the observer reflects the f - u or g - u curves of the gravitational wave after suitable rescaling of the horizontal and vertical coordinates. The two basic polarization modes of a gravitational wave are: ① $G = 0$ [and thus $g(u) \equiv 0$], corresponding to mode $+$; ② $F = 0$ [and thus $f(u) \equiv 0$], corresponding to mode \times . In the approximation where the gravitational wave is weak enough, the oscillation patterns of the circle in one period under the actions of these two modes are illustrated in Fig. 7.23. A general oscillation can be expressed as the superposition of these two modes, which has been discussed in Sect. 7.9.2.

The effect of a gravitational wave on the test particles shown in Figs. 7.22 and 7.23 is different from that of an electromagnetic wave. A gravitational wave is the “propagation of the oscillation of curvature”, and curvature leads to tidal acceleration. Therefore, the

Fig. 7.22 The deformation of a circle under a gravitational wave at a certain time (see Table 7.2)

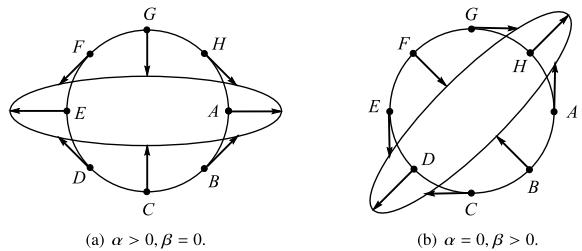
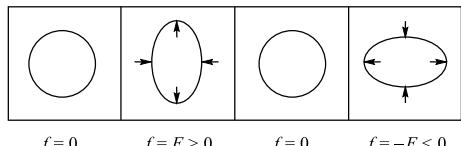


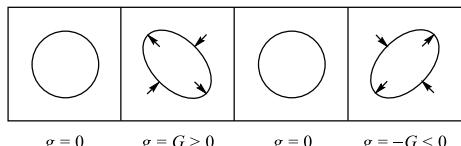
Table 7.2 The tidal acceleration \vec{a} of 8 points on a circle relative to the center (see Fig. 7.22 for the overall effect)

(a) $\alpha = 0, \beta > 0$	A	B	C	D	E	F	G	H
$\begin{bmatrix} w^1 \\ w^2 \\ w^3 \\ w^4 \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ 0 & -\alpha \end{bmatrix} \begin{bmatrix} w^1 \\ w^2 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \\ \alpha \end{bmatrix}$	$\begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} -1/\sqrt{2} \\ -1/\sqrt{2} \\ -\alpha/\sqrt{2} \\ \alpha/\sqrt{2} \end{bmatrix}$	$\begin{bmatrix} -1 \\ 0 \\ -\alpha \\ 0 \end{bmatrix}$	$\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ -\alpha/\sqrt{2} \\ -\alpha/\sqrt{2} \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \\ 0 \\ -\alpha \end{bmatrix}$	$\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ \alpha/\sqrt{2} \\ -\alpha/\sqrt{2} \end{bmatrix}$
\vec{a}	\bullet_A	\bullet_B	\bullet_C	\bullet_D	\bullet_E	\bullet_F	\bullet_G	\bullet_H
(b) $\alpha = 0, \beta > 0$	A	B	C	D	E	F	G	H
$\begin{bmatrix} w^1 \\ w^2 \\ w^3 \\ w^4 \end{bmatrix} = \begin{bmatrix} 0 & \beta \\ \beta & 0 \end{bmatrix} \begin{bmatrix} w^1 \\ w^2 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ \beta \end{bmatrix}$	$\begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ -\beta/\sqrt{2} \\ \beta/\sqrt{2} \end{bmatrix}$	$\begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} -1/\sqrt{2} \\ -1/\sqrt{2} \\ -\beta/\sqrt{2} \\ -\beta/\sqrt{2} \end{bmatrix}$	$\begin{bmatrix} -1 \\ 0 \\ 0 \\ -\beta \end{bmatrix}$	$\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ \beta/\sqrt{2} \\ -\beta/\sqrt{2} \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \\ \beta \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ \beta/\sqrt{2} \\ \beta/\sqrt{2} \end{bmatrix}$
\vec{a}	\bullet_A	\bullet_B	\bullet_C	\bullet_D	\bullet_E	\bullet_F	\bullet_G	\bullet_H

Fig. 7.23 The oscillation of a circle in one period under a linearly polarized gravitational plane wave

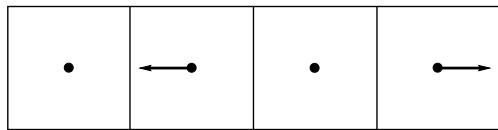


(a) Polarization mode $+ [G = 0, F > 0, f = F \cos(\omega t - kz)]$.



(b) Polarization mode $\times [G > 0, F = 0, g = G \cos(\omega t - kz)]$.

Fig. 7.24 The oscillation of a charged particle in one period under a linearly polarized electromagnetic wave



gravitational wave can be detected by measuring the relative acceleration between a free particle and another (fiducial) free particle, as we have discussed above. An electromagnetic wave is the propagation of the oscillation of the electromagnetic field, when detecting one just needs to measure the acceleration of a charged particle relative to an inertial frame, whose expression, namely $\ddot{a} = (q/m)\vec{E}$, is much simpler than the tidal acceleration. Suppose the electromagnetic wave being detected is linearly polarized, then what corresponds to Fig. 7.23 is the much simpler Fig. 7.24. We have shown in Optional Reading 7.9.2 that the polarization tensor of a gravitational wave will come back to itself after rotating by (an integer times) π about the z -axis, while the polarization vector of an electromagnetic wave will come back after rotating by (an integer times) 2π . This difference is also manifested by the polarization patterns in Figs. 7.23 and 7.24: the pattern in any square in Fig. 7.23 (i.e., at any time) will come back to itself after rotating by (an integer times) π about the direction of propagation (the line perpendicular to the page that passes through the centre of symmetry), while the pattern in any square in Fig. 7.24 will come back after rotating by (an integer times) 2π . This difference between Figs. 7.23 and 7.24 reflects again the fact that photons are spin-1 while gravitons are spin-2, as mentioned in Optional Reading 7.9.2.

[The End of Optional Reading 7.9.5]

Exercises

- ~7.1. Show that Maxwell's equation in curved spacetime, $\nabla^a F_{ab} = -4\pi J_b$, contains the law of conservation of charge, i.e., $\nabla_a J^a = 0$. NB: $\nabla^a F_{ab} = -4\pi J_b$ is equivalent to (7.2.8) rather than (7.2.9), and hence this problem indicates that (7.2.8) rather than (7.2.9) gives the charge conservation.
- ~7.2. Show that $\frac{D\omega_a}{d\tau} = \frac{D\omega_a}{d\tau} + (A_a \wedge Z_b)\omega^b$, $\forall \omega_a \in \mathcal{F}_G(0, 1)$.

- ~7.3. Prove property (3) of the Fermi derivative in Proposition 7.3.1.
- 7.4. Show that a (nonvanishing) vector field v^a on a timelike curve $G(\tau)$ with a constant magnitude must undergo a spacetime rotation. Hint: Let $u^a \equiv Dv^a/d\tau$, then $u_a v^a = 0$. First show that no matter whether $v_a v^a$ vanishes or not, one always has a vector field v'^a on $G(\tau)$ such that $v'_a v^a = 1$. And then show that v^a undergoes a spacetime rotation with the angular velocity 2-form $\Omega_{ab} \equiv 2v'_{[a} u_{b]}$.
- 7.5. Suppose $\{T, X, Y, Z\}$ is a Lorentzian coordinate system in Minkowski space-time, and the parametric equation of a curve $G(\tau)$ is

$$T = A^{-1} \sinh A\tau, \quad X = A^{-1} \cosh A\tau, \quad Y = Z = 0 \quad (A \text{ is a constant}).$$

- (a) Show that $G(\tau)$ is a timelike hyperbola (i.e., G in Fig. 6.43), τ is the proper time, and A is the magnitude of the 4-acceleration A^a of $G(\tau)$.
- *(b) Show that any ray $\mu(s)$ starting from the origin o of the system $\{T, X, Y, Z\}$ that intersects $G(\tau)$ is orthogonal to $G(\tau)$.
- *(c) Suppose the parameter s of $\mu(s)$ in (b) is the arc length of μ , as we collect all of the rays $\mu(s)$ starting from o that intersect $G(\tau)$, we obtain a spatial vector field $w^a \equiv (\partial/\partial s)^a$ on $G(\tau)$. Show that w^a is Fermi transported along $G(\tau)$.
- *(d) Let $Z^a \equiv (\partial/\partial \tau)^a$, and choose $\{Z^a, w^a, (\partial/\partial Y)^a, (\partial/\partial Z)^a\}$ as an orthonormal tetrad field on $G(\tau)$, find the proper coordinate system $\{t, x, y, z\}$ of $G(\tau)$ and specify its coordinate patch.
- Answer: $T = (A^{-1} + x) \sinh At$, $X = (A^{-1} + x) \cosh At$, $Y = y$, $Z = z$.
- (e) Write down the expression for the line element of the Minkowski metric in the above proper coordinate system. Compute the Christoffel symbol of the Minkowski metric in this system, and verify that it satisfies Lemma 7.4.3, i.e., (7.4.10).
- 7.6. Suppose G is a non-rotating, freely falling, instantaneous rest observer of a point mass L at a point $p \in L$ (i.e., the 4-velocity Z^a of G and the 4-velocity U^a of L are tangent at p), A^a is the 4-acceleration of L at p , and a^a is the 3-acceleration of L at p relative to G [defined by (7.4.3)]. Show that $a^a = A^a$. NB: This claim can be viewed as the generalization of Proposition 6.3.6 to curved spacetime.
- ~7.7. A metric g_{ab} is said to be **Ricci flat** if the Ricci tensor of g_{ab} vanishes. Show that a necessary and sufficient condition for a 4-dimensional Lorentzian metric g_{ab} being a solution to the vacuum Einstein equation is that g_{ab} is Ricci flat.
- ~7.8. Suppose (M, g_{ab}) is a Ricci flat spacetime (see the above problem for the definition), and ξ^a is one of the Killing vector fields of the spacetime. Show that $F_{ab} := (d\xi)_{ab}$ satisfies the source-free ($J_a = 0$) Maxwell equation of (M, g_{ab}) . Hint: use $\nabla_a \xi^a = 0$ satisfied by any Killing vector field ξ^a (the result of Exercise 4.11).
- 7.9. Suppose γ_{ab} satisfies (a) $\partial^a \bar{\gamma}_{ab} = 0$; (b) $\gamma = 0$; (c) $\gamma_{0i} = 0$ ($i = 1, 2, 3$); (d) $\gamma_{00} = \text{constant}$. Find an “infinitesimal” vector field ξ^a such that $\tilde{\gamma}_{ab} \equiv \gamma_{ab} + \partial_a \xi_b + \partial_b \xi_a$ satisfies the transverse-traceless gauge conditions:
 (a) $\partial^a \tilde{\gamma}_{ab} = 0$; (b) $\tilde{\gamma} = 0$; (c) $\tilde{\gamma}_{0i} = 0$ ($i = 1, 2, 3$); (d) $\tilde{\gamma}_{00} = 0$.

- 7.10. Suppose $g_{ab} = \eta_{ab} + \gamma_{ab}$ represents a gravitational wave with γ_{ab} of the form (7.9.30), $\{t, x^i\}$ is a Lorentzian coordinate system, ∇_a is the derivative operator associated with g_{ab} , and $Z^a \equiv (\partial/\partial t)^a$. Show that $Z^a \nabla_a Z^b = 0$, i.e., the t -coordinate lines are geodesics. Hint: compute $\mathcal{L}_Z g_{ab}$ by plugging in the ansatz (7.9.30), contract it with Z^a , then use (4.3.1') and the TT gauge condition. NB: By a similar proof, this conclusion can also be applied to the elliptically polarized waves of the form (7.9.63).
- 7.11. Prove Proposition 7.9.5.
- 7.12. Verify the properties ①–③ of $\{E_i\}^a$ in (7.9.84).
- 7.13. Prove (7.9.86).
- 7.14. Prove (7.9.78), i.e., $\nabla^a \nabla_a P = (\partial^2 P / \partial x^2) + (\partial^2 P / \partial y^2)$.

References

- Abbott, B. P. et al. (2016), ‘Observation of Gravitational Waves from a Binary Black Hole Merger’, *Phys. Rev. Lett.* **116**(6), 061102. [arXiv:1602.03837](#).
- Cai, R.-G., Cao, Z., Guo, Z.-K., Wang, S.-J. and Yang, T. (2017), ‘The Gravitational-Wave Physics’, *Natl. Sci. Rev.* **4**(5), 687–706. [arXiv:1703.00187](#).
- Carroll, S. M. (2019), *Spacetime and Geometry*, Cambridge University Press, Cambridge.
- Chen, C.-M., Nester, J. M. and Ni, W.-T. (2017), ‘A brief history of gravitational wave research’, *Chin. J. Phys.* **55**, 142–169. [arXiv:1610.08803](#).
- Fock, V. A. (1939), ‘Sur le mouvement des masses finies d’Apres la theorie de gravitation Einsteinienne’, *J. Phys. U.S.S.R.* **1**, 81–166.
- Geroch, R. P. and Jang, P. S. (1975), ‘Motion of a body in general relativity’, *J. Math. Phys.* **16**, 65–67.
- Hawking, S. W. and Ellis, G. F. R. (1973), *The Large Scale Structure of Space-Time*, Cambridge University Press, Cambridge.
- d’Inverno, R. A. (1992), *Introducing Einstein’s Relativity*, Clarendon Press, Oxford.
- Liu, L. and Zhao, Z. (2004), *General Relativity (in Chinese)*, Higher Education Press, Beijing.
- Maggiore, M. (2018), *Gravitational Waves: Volume 2: Astrophysics and Cosmology*, Oxford University Press, Oxford.
- Misner, C., Thorne, K. and Wheeler, J. (1973), *Gravitation*, W H Freeman and Company, San Francisco.
- Ohanian, H. C. and Ruffini, R. (1994), *Gravitation and Spacetime*, W W Norton and Company, Inc., New York.
- Sachs, R. K. and Wu, H. (1977), *General Relativity for Mathematicians*, Springer-Verlag, New York.
- Saulson, P. R. (2017), *Fundamentals Of Interferometric Gravitational Wave Detectors*, World Scientific, Singapore.
- Stephani, H., Kramer, D., MacCallum, M. A. H., Hoenselaers, C. and Herlt, E. (2003), *Exact Solutions of Einstein’s Field Equations*, Cambridge University Press, Cambridge.
- Straumann, N. (1984), *General Relativity and Relativistic Astrophysics*, Springer-Verlag, Berlin.
- Synge, J. L. (1960), *Relativity: The General Theory*, North-Holland Publishing Company, Amsterdam.
- Wald, R. M. (1984), *General Relativity*, The University of Chicago Press, Chicago.
- Weber, J. (1961), *General Relativity and Gravitational Waves*, Wiley-Interscience, New York.

- Will, C. M. (1995), Stable clocks and general relativity, in ‘30th Rencontres de Moriond: Euro-conferences: Dark Matter in Cosmology, Clocks and Tests of Fundamental Laws’, pp. 417–428. [arXiv:gr-qc/9504017](https://arxiv.org/abs/gr-qc/9504017).
- Will, C. M. (2014), ‘The confrontation between general relativity and experiment’, *Living Reviews in Relativity* **17**(1), 4. [arXiv:1403.7377](https://arxiv.org/abs/1403.7377).
- Will, C. M. (2018), *Theory and Experiment in Gravitational Physics*, Cambridge University Press, Cambridge.

Chapter 8

Solving Einstein's Equation



Solving Einstein's Equation is an important problem in general relativity. Many exact solutions play important roles in the study and development of general relativity. Since Einstein's equation is a highly nonlinear partial differential equation, finding an (exact) solution in the general case is rather difficult. The first exact solution—the vacuum Schwarzschild solution—was found by Karl Schwarzschild under the premise that the spacetime is static and has spherical symmetry. Schwarzschild's solution, which is often regarded as one of the most important solutions in general relativity, was found within two months after Einstein's equation was published.¹

8.1 Stationary Spacetimes and Static Spacetimes

Definition 1 A spacetime (M, g_{ab}) is said to be **stationary** if it has a timelike Killing vector field. In this case, we also call g_{ab} a stationary metric.

Suppose there exists a timelike Killing vector field ξ^a in (M, g_{ab}) , whose integral curves have the parameter t , i.e., $\xi^a = (\partial/\partial t)^a$. Choose any coordinate system $\{x^\mu\}$ where t is the zeroth coordinate (i.e., $t = x^0$) and the integral curve of ξ^a is the x^0 -coordinate line (namely the coordinate system adapted to ξ^a , see Sect. 4.2). Let $g_{\mu\nu}$ be the components of g_{ab} in this coordinate system, then

$$\frac{\partial g_{\mu\nu}}{\partial t} = (\mathcal{L}_\xi g)_{\mu\nu} = 0, \quad (8.1.1)$$

¹ To be precise, within 34 days. Inspired by Einstein's Mercury perihelion result of November 18, 1915, he looked for an exact solution. He communicated what he found in a letter to Einstein on December 22, 1915. His solution was published in January 1916. Furthermore, this was in the middle of World War I, and Schwarzschild was in the army on the Russian front!

where we used Theorem 4.2.2 in the first equality, and the second equality is due to the fact that ξ^a is a Killing vector field. Equation (8.1.1) indicates that all of the components $g_{\mu\nu}$ are independent of the time coordinate t , i.e., $g_{\mu\nu}$ is “time-translation invariant”. This is exactly where the term “stationary” comes from.

Inversely, if there exists a local coordinate system $\{x^\mu\}$ in (M, g_{ab}) such that

$$\frac{\partial g_{\mu\nu}}{\partial t} = 0 \quad (t \equiv x^0 \text{ is a timelike coordinate}), \quad (8.1.2)$$

then $\xi^a \equiv (\partial/\partial t)^a$ is a smooth vector field on the coordinate patch O , and $\{x^\mu\}$ is exactly a coordinate system adapted to this vector field. Hence, it follows from Theorem 4.2.2 that

$$(\mathcal{L}_\xi g)_{\mu\nu} = \frac{\partial g_{\mu\nu}}{\partial t} = 0,$$

which means that on O we have $\mathcal{L}_\xi g_{ab} = 0$, and thus $\xi^a \equiv (\partial/\partial t)^a$ is a timelike Killing vector field. Therefore, a stationary space can also be defined in terms of the coordinate language as follows: if there exists a local coordinate system $\{x^\mu\}$ (whose coordinate patch is O) such that all of the components of g_{ab} are independent of the timelike coordinate x^0 , then (O, g_{ab}) is a stationary spacetime.

Intuitively speaking, a stationary spacetime corresponds to a gravitational field that does not change with time. However, the notion of time depends on the observer. For instance, since the Earth's gravitational field on the ground is stronger than that in the upper air, you (as an observer) will find that the Earth's gravitational field “changes with time” if you keep measuring the gravitational field while moving from the ground up into the air. This certainly does not indicate that the Earth's gravitational field is not a stationary gravitational field. Thus, when judging the stationarity of a gravitational field by means of an observer, one needs to choose an appropriate observer (reference frame). If you somehow can keep yourself at a fixed height above a certain point on the ground (your world line is parallel to a generatrix of the world sheet of the Earth's surface), you will see that the Earth's gravitational field “does not change with time”. That is, the spacetime corresponding to the Earth's gravitational field has the following property: there exists a specific class of timelike curves (which coincides with the integral curves of the timelike Killing vector field), such that the metric components measured by the observers whose world lines are these curves do not change with time. Many spacetimes (e.g., the expanding universe) do not have this property (i.e., do not have a timelike Killing vector field), and Definition 1 is exactly the mathematical formulation of this property.

Example 1 Minkowski spacetime is a stationary spacetime, since the zeroth coordinate basis vector field $(\partial/\partial x^0)^a$ of its Lorentzian coordinate system $\{x^\mu\}$ is a timelike Killing vector field.

Example 2 The metric of a certain 2-dimensional spacetime can be expressed in some coordinate system $\{t, x\}$ as $ds^2 = -t^{-4}dt^2 + dx^2$, $t > 0$. Some people may say this is not a stationary metric since its component $g_{00} = -t^{-4}$ depends on the

time coordinate t . However, a simple coordinate transformation $T = t^{-1}$, $X = x$ will turn the line element into $ds^2 = -dT^2 + dX^2$. This is nothing but a 2-dimensional Minkowski metric, which is of course stationary!

Example 2 in a way suggests that confusion may arise if one does not take the geometric perspective. Stationarity is an intrinsic property of the spacetime geometry, which does not depend on the choice of the coordinate system. Note that both of the following statements are wrong:

(1) (WRONG!) If some coordinate components $g_{\mu\nu}$ of the metric depend on the timelike coordinate t of this coordinate system, then the spacetime is not stationary.

(2) (WRONG!) The spacetime in Example 2 is a stationary spacetime in the coordinate system $\{T, X\}$, but is not a stationary spacetime in the coordinate system $\{t, x\}$.

Definition 2 A vector field v^a in (M, g_{ab}) is said to be **hypersurface orthogonal** if $\forall p \in M$ there exists a hypersurface Σ that is everywhere orthogonal to v^a such that $p \in \Sigma$.

Definition 3 A spacetime (M, g_{ab}) is said to be **static** if it has a hypersurface orthogonal timelike Killing vector field. In this case, we also call g_{ab} a **static metric**.

Thus, a static spacetime must be stationary, but not vice versa.

Proposition 8.1.1 Suppose $\xi^a = (\partial/\partial t)^a$ is a Killing vector field, and $\Sigma_0 = \{p \in M | t(p) = 0\}$ is a hypersurface everywhere orthogonal to ξ^a , then the hypersurface $\Sigma_{t_1} = \{p \in M | t(p) = t_1\}$ is also everywhere orthogonal to ξ^a .

Proof Exercise 8.1. Hint: $\Sigma_t = \phi_t[\Sigma_0]$, where ϕ_t is an element in the one-parameter group of isometries corresponding to ξ^a , i.e., an isometry. \square

Suppose (M, g_{ab}) is a static spacetime, $\xi^a = (\partial/\partial t)^a$ is a timelike Killing field, and Σ_0 is a hypersurface orthogonal to ξ^a . Choose the intersection of Σ_0 and each integral curve of ξ^a as the zero of the curve's parameter, and choose a local coordinate system $\{x^i\}$ on Σ_0 . Since we have $\xi^a \neq 0$ at each point on Σ_0 , we can “carry” these three coordinates outside Σ_0 (i.e., set the x^i of each point on the integral curve of ξ^a to be the x^i of the intersection of Σ and this curve), and take the parameter t of each integral curve as the timelike coordinate x^0 (called the Killing coordinate time) of each point on the curve, then we obtain a 4-dimensional local coordinate system $\{t, x^i\}$, whose t -coordinate lines are the integral curves of ξ^a . Also, since the x^i -coordinate lines lie on the orthogonal surface Σ_t , the timelike coordinate basis vector $(\partial/\partial t)^a$ is orthogonal to the spacelike coordinate basis vectors $(\partial/\partial x^i)^a$. Therefore,

$$g_{0i} = g_{ab}(\partial/\partial t)^a(\partial/\partial x^i)^b = 0, \quad i = 1, 2, 3,$$

and hence the expression for the line element of g_{ab} in this system is simplified as

$$ds^2 = g_{00}(x^1, x^2, x^3)dt^2 + g_{ij}(x^1, x^2, x^3)dx^i dx^j. \quad (8.1.3)$$

Such a coordinate system is called a **time-orthogonal coordinate system**.

Suppose (M, g_{ab}) is a stationary spacetime, then the reference frame corresponding to the integral curves of the timelike Killing vector field ξ^a is called a **stationary reference frame** (“corresponding to” means to reparametrize the integral curves and substitute the Killing time t with the proper time τ). A stationary reference frame whose ξ^a is hypersurface orthogonal is called a **static reference frame**. An observer is called a **stationary (static) observer** if they are an observer of a stationary (static) reference frame. The Σ_t defined in Proposition 8.1.1 is called a **surface of simultaneity** of the static reference frame. Note that the “time” t here is the coordinate time rather than the proper time τ of the static observer (unless $g_{00} = -1$); it is easy to show that they have the following relation: $d\tau = \sqrt{-g_{00}}dt$.

A static spacetime has not only a time-translation invariance that any stationary spacetime has, but also a time-reflection invariance (except for some possible subtle cases). Suppose $\xi^a = (\partial/\partial t)^a$ is a hypersurface orthogonal timelike Killing vector field, then a time reflection transformation is referring to the diffeomorphism $\phi : M \rightarrow M$ satisfying $t(\phi(p)) = -t(p)$, $x^i(\phi(p)) = x^i(p)$, $\forall p \in M$. Now we will show that this ϕ is an isometry, and thus a static spacetime is said to be time reflection invariant.

Suppose $C(t)$ is the integral curve of ξ^a passing through p , and $p = C(t_1)$. From $x^i(\phi(p)) = x^i(p)$ we can see that $q \equiv \phi(p)$ is also on $C(t)$. First we show that $\phi_*[(\partial/\partial t)^a|_p] = -(\partial/\partial t)^a|_q$. Let $v^a \equiv (\partial/\partial t)^a|_p$, $u^a \equiv -(\partial/\partial t)^a|_q$, $r \equiv C(t_1 + \Delta t)$, and $s \equiv \phi(r)$ (see Fig. 8.1). Suppose f is an arbitrary smooth function on M , then the result of the vector ϕ_*v^a at q acting on f is

$$\begin{aligned} (\phi_*v)(f) &= v(\phi^*f) = \frac{\partial}{\partial t} \Big|_{t=t_1} (\phi^*f) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} [(\phi^*f)|_r - (\phi^*f)|_p] \\ &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} (f|_s - f|_q) = u(f), \end{aligned}$$

and hence $\phi_*v^a = u^a$, i.e., $\phi_*[(\partial/\partial t)^a|_p] = -(\partial/\partial t)^a|_q$. Similarly, one can show that

$$\phi_*[(\partial/\partial x^i)^a|_p] = (\partial/\partial x^i)^a|_q, \quad i = 1, 2, 3.$$

Let $g_{\mu\nu}$ and $(\phi^*g)_{\mu\nu}$ represent the components of g_{ab} and $(\phi^*g)_{ab}$, respectively, in the system $\{t, x^i\}$, then

Fig. 8.1 Time reflection
 $\phi : M \rightarrow M$

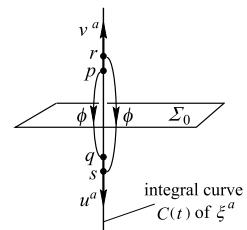
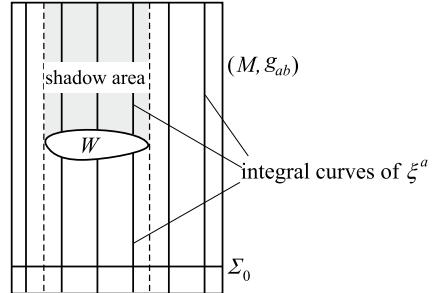


Fig. 8.2 A strong static spacetime becomes a weak static spacetime after W is removed



$$\begin{aligned}
 (\phi^*g)_{00}|_p &= [(\phi^*g)_{ab}(\partial/\partial t)^a(\partial/\partial t)^b]|_p \\
 &= [g_{ab}(\phi_*\partial/\partial t)^a(\phi_*\partial/\partial t)^b]|_q \\
 &= [g_{ab}(\partial/\partial t)^a(\partial/\partial t)^b]|_q = g_{00}|_q = g_{00}|_p,
 \end{aligned}$$

where the last step is because of $0 = (\mathcal{L}_\xi g)_{\mu\nu} = \partial g_{\mu\nu}/\partial t$, i.e., $g_{\mu\nu}$ are constants along $C(t)$. Similarly, we have $(\phi^*g)_{ij}|_p = g_{ij}|_p$, but $(\phi^*g)_{0i}|_p = -g_{0i}|_p$. Luckily, $g_{0i} = 0$ (where the hypersurface orthogonality is used), and hence $(\phi^*g)_{\mu\nu}|_p = g_{\mu\nu}|_p$. Noticing that p is arbitrary, we know that $(\phi^*g)_{ab} = g_{ab}$, and so $\phi : M \rightarrow M$ is an isometry.

[Optional Reading 8.1.1]

Technically, the definition of a Killing vector field has a strong version and a weak version. The weak definition only cares about the local properties: any vector field ξ^a satisfying the Killing equation $\nabla_{(a}\xi_{b)} = 0$ (equivalent to $\mathcal{L}_\xi g_{ab} = 0$) is called a Killing vector field. This ξ^a may be incomplete, i.e., the range of its parameter t is not the whole \mathbb{R} but an interval of \mathbb{R} . The strong definition, however, requires that ξ^a be complete. Accordingly, the definitions of stationary and static spacetimes also have a weak version and a strong one, depending on whether or not the timelike Killing vector field is complete. When we are only concerned with local issues, it is not necessary to emphasize the difference between them; however, when global issues are involved, some conclusions only hold if the spacetime satisfies the strong condition. For instance, if a region W is removed from a strong static spacetime (M, g_{ab}) , this spacetime will become a weak static spacetime. Suppose what is shown in Fig. 8.2 is the Σ_0 in Proposition 8.1.1, then $\Sigma_{t_1} = \{p \in M | t(p) = t_1\}$ is meaningless when t is sufficiently large, since the Killing field ξ^a is not well-defined at the zero of the parameter t of each integral curve inside the “shadow region” (and hence t is not well-defined). Thus, it is possible that Proposition 8.1.1 only holds locally for a static spacetime.

In a word, the key difference between the strong and weak definitions is whether ξ^a is complete or not. ξ^a generates a one-parameter group of isometries when it is complete, while it only generates a one-parameter local group of isometries when it is incomplete. For convenience’s sake, we usually omit the word “local” in the text.

[The End of Optional Reading 8.1.1]

8.2 Spherically Symmetric Spacetimes

First we discuss a 2-dimensional sphere (S^2, h_{ab}) in the 3-dimensional Euclidean space ($\mathbb{R}^3, \delta_{ab}$), where h_{ab} is the induced metric of δ_{ab} . The expression for the line element of h_{ab} in the spherical coordinate system $\{\theta, \varphi\}$ is

$$ds^2 = r^2(d\theta^2 + \sin^2 \theta d\varphi^2),$$

where r is the radius of the sphere. Without loss of generality, here we only talk about the unit sphere ($r = 1$), whose line element is

$$ds^2 = d\theta^2 + \sin^2 \theta d\varphi^2. \quad (8.2.1)$$

It follows from the equation above that

$$\xi_1^a \equiv (\partial/\partial\varphi)^a \quad (8.2.2a)$$

is a Killing vector field, which reflects the invariance of (S^2, h_{ab}) under a rotation with respect to the z -axis. The integral curves of such rotations are all the circles of latitude on the sphere (the circle at each of the two poles shrinks to a point), see Fig. 8.3. It is intuitively not difficult to believe that (S^2, h_{ab}) has maximal symmetry, and thus should have 3 independent Killing vector fields. In fact, it does. It is not difficult to verify that

$$\xi_2^a \equiv (\partial/\partial\theta)^a \sin \varphi + (\partial/\partial\varphi)^a \cot \theta \cos \varphi, \quad (8.2.2b)$$

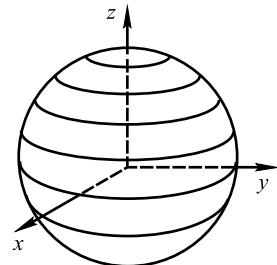
and

$$\xi_3^a \equiv [\xi_1, \xi_2]^a = (\partial/\partial\theta)^a \cos \varphi - (\partial/\partial\varphi)^a \cot \theta \sin \varphi \quad (8.2.2c)$$

are also Killing fields, and $\xi_1^a, \xi_2^a, \xi_3^a$ are linearly independent. From Sect. 4.3 we have learned that the one-parameter group of diffeomorphisms corresponding to a Killing vector field is a one-parameter group of isometries, and hence the collection of all the isometries on (S^2, h_{ab}) is a 3-parameter group, which is isomorphic to the rotation group $SO(3)$ of the 3-dimensional Euclidean space. Readers who are not familiar with group theory do not have to worry too much about this, one just needs to know that $SO(3)$ is such a group, each element of which is a rotation that keeps the origin in the 3-dimensional Euclidean space fixed (see Appendix G in Volume II for details).

When talking about spacetime symmetries, one should pay attention to the relation and difference between isometries and diffeomorphisms. An isometry must be a diffeomorphism, but not vice versa. Each smooth vector field corresponds to a one-parameter group of diffeomorphisms (we will omit the term “local” from now on), and so any manifold M has infinitely many one-parameter groups of diffeomorphisms. The collection of all the diffeomorphisms is a group of infinitely many

Fig. 8.3 The integral curves of a Killing vector field on a sphere



parameters, called the **diffeomorphism group** on M . Each Killing vector field on (M, g_{ab}) corresponds to a one-parameter group of isometries, which is a subgroup of the diffeomorphism group on M . The collection of all the isometries is called the **isometry group** of (M, g_{ab}) . Since a 4-dimensional spacetime has at most 10 independent Killing vector fields, the isometry group of this spacetime has at most 10 parameters. Suppose G_1 is a one-parameter group of diffeomorphisms on M , then $\forall p \in M$, the collection of points obtained by acting each element of G_1 on p is called an orbit of G_1 passing through p (see Sect. 2.2). This definition of an orbit can be carried over to any subgroup of the diffeomorphism group on M . It is not difficult to see the following: suppose G_3 is the isometry group on (S^2, g_{ab}) [which is isomorphic to $SO(3)$], then any orbit of G_3 passing through $p \in S^2$ is S^2 itself.

Definition 1 A spacetime (M, g_{ab}) is said to be spherically symmetric if its isometry group has a subgroup G_3 that is isomorphic to $SO(3)$ and all the orbits of G_3 (except for the fixed points) are 2-dimensional spheres. These spheres are called **orbit spheres**.

Remark 1 ① The isometry group of a spherically symmetric spacetime can be larger than $SO(3)$. For instance, the isometry group of Minkowski spacetime has 10 parameters, but it is a spherically symmetric spacetime, since it contains a subgroup isomorphic to $SO(3)$, whose orbits (except for a fixed point) are all 2-dimensional spheres. ② Precisely speaking, Definition 1 only defines a **spherically symmetric metric field** rather than a spherically symmetric spacetime. If there exists a matter field in spacetime (i.e., $T_{ab} \neq 0$), then (M, g_{ab}) is called a **spherically symmetric spacetime** only if the metric field and the matter field are both spherically symmetric (Sect. 8.6 will involve the relation between the symmetry of the matter field and the symmetry of the metric field).

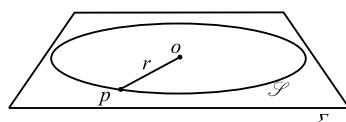


Fig. 8.4 An orbit sphere \mathcal{S} on a surface Σ of simultaneity of an inertial frame in Minkowski spacetime (with one dimension suppressed)

The subgroup G_3 of the isometry group which is isometric to $SO(3)$ corresponds to three independent Killing vector fields ξ_1^a, ξ_2^a , and ξ_3^a . Suppose \mathcal{S} is an orbit of G_3 (a 2-sphere), then the integral curves of $\xi_1^a, \xi_2^a, \xi_3^a$ starting from any point on \mathcal{S} all lie on \mathcal{S} , and hence $\xi_1^a, \xi_2^a, \xi_3^a$ at any point on \mathcal{S} are all tangent to \mathcal{S} . Suppose \hat{g}_{ab} is the 2-dimensional metric on \mathcal{S} induced by g_{ab} , then from the definition of an induced metric we can see that $\xi_1^a, \xi_2^a, \xi_3^a$ on \mathcal{S} are also Killing fields measured by \hat{g}_{ab} , and thus $(\mathcal{S}, \hat{g}_{ab})$ has the maximal symmetry represented by ξ_1^a, ξ_2^a , and ξ_3^a . Therefore (see Optional Reading 8.2.1 for a proof), \hat{g}_{ab} can only be a standard spherical metric h_{ab} (the metric induced on a sphere by the 3-dimensional Euclidean metric), i.e., there exists a constant $K > 0$ and a coordinate system $\{\theta, \varphi\}$ such that the line element of \hat{g}_{ab} can be expressed by

$$d\hat{s}^2 = K(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (8.2.3)$$

Take Minkowski spacetime as an example. Suppose Σ is a surface of simultaneity of an inertial frame. By assigning a set of concentric 2-spheres on Σ (see Fig. 8.4), we can pick out from the 10-dimensional isometry group a subgroup G_3 isomorphic to $SO(3)$, whose orbit passing through any point p in Σ (except for the center o) is the sphere that p lives in. The line element of the Minkowski metric in the chosen inertial coordinate system is

$$ds^2 = -dt^2 + dr^2 + d\hat{s}^2,$$

where

$$d\hat{s}^2 = r^2(d\theta^2 + \sin^2 \theta d\varphi^2).$$

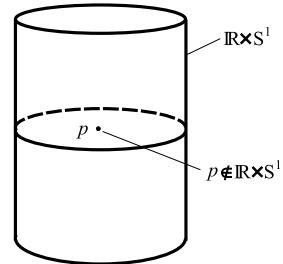
Thus, for the Minkowski metric, the K in (8.2.3) is the square of the radius of the orbit 2-sphere \mathcal{S} which we have been discussing. To figure out the meaning of K in a non-flat spacetime, a geometric concept will be helpful to us, namely the area of \mathcal{S} . Suppose $\hat{\epsilon}$ is the area element on \mathcal{S} associated with \hat{g}_{ab} , then the area of \mathcal{S} will be $A = \int_{\mathcal{S}} \hat{\epsilon}$. Also, $\hat{\epsilon}$ can be expressed using the coordinate system $\{\theta, \varphi\}$ on \mathcal{S} as $\hat{\epsilon} = \sqrt{\hat{g}} d\theta \wedge d\varphi$, in which \hat{g} is the determinant of \hat{g}_{ab} in the system $\{\theta, \varphi\}$. After reading off \hat{g}_{ij} from (8.2.3) we can find $\hat{g} = K^2 \sin^2 \theta$, and hence $\hat{\epsilon} = K \sin \theta d\theta \wedge d\varphi$. Therefore,

$$A = K \int_0^{2\pi} d\varphi \int_0^\pi \sin \theta d\theta = 4\pi K.$$

Thus, K is the area of the sphere divided by 4π . Define

$$r := \sqrt{\frac{A}{4\pi}}, \quad (8.2.4)$$

Fig. 8.5 A cylindrical surfaces in the 3-dimensional Euclidean space. The center p of any circle in the surface is not on the surface



and call r the **radius**, then $K = r^2$, and (8.2.3) can be rewritten as

$$d\hat{s}^2 = r^2(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (8.2.5)$$

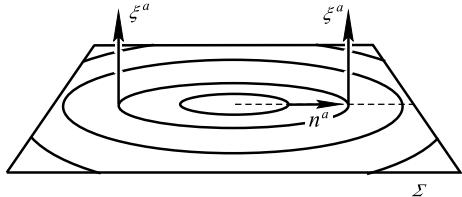
Seemingly, this is the same as the expression (8.2.3) for the $d\hat{s}^2$ of Minkowski spacetime, and the r in each equation is called the radius. However, the radius r in general does not necessarily have the meaning of “the distance between the center and each point on \mathcal{S} ”. In fact, the following three cases are all possible: ① There does not exist a point that can be regarded as the center of \mathcal{S} at all. Let us look at a simplified example: suppose S^1 is a circle in the manifold $\mathbb{R} \times S^1$ (a cylindrical surface), then the center of S^1 will not be on the manifold $\mathbb{R} \times S^1$ (Fig. 8.5). Similarly, in $\mathbb{R} \times S^2$ there does not exist a point that can be regarded as the center of S^2 either. ② There exists a point in the spacetime that can be regarded as the center of \mathcal{S} , but due to the curved metric, the distance between \mathcal{S} and this point is not equal to the radius r defined by (8.2.4). ③ There exists more than one center of \mathcal{S} .

[Optional Reading 8.2.1]

Before we wrote down (8.2.3) we have assumed the following claim: suppose $(\mathcal{S}, \hat{g}_{ab})$ has the maximal symmetry represented by ξ_1^a , ξ_2^a , and ξ_3^a , then the line element of \hat{g}_{ab} can always be expressed as (8.2.3). Now we briefly introduce how to prove this claim. Suppose the components of \hat{g}_{ab} in the coordinate system $\{\theta, \varphi\}$ are \hat{g}_{11} , \hat{g}_{22} and \hat{g}_{12} , then from $\xi_1^a = (\partial/\partial\varphi)$ we can see that \hat{g}_{11} , \hat{g}_{22} and \hat{g}_{12} are not functions of φ . Writing down the equations of the coordinate components of $\mathcal{L}_{\xi_2} \hat{g}_{ab} = 0$ satisfied by ξ_2 , and taking $\hat{g}_{11}(\theta)$, $\hat{g}_{22}(\theta)$ and $\hat{g}_{12}(\theta)$ as functions to be solved for, we obtain $\hat{g}_{12} = 0$, $\hat{g}_{11} = K$ (constant) and $\hat{g}_{22} = K \sin^2 \theta$. It is not difficult to verify that $\mathcal{L}_{\xi_3} \hat{g}_{ab} = 0$, which completes the proof.

[The End of Optional Reading 8.2.1]

Fig. 8.6 The orbit sphere passing through any point on Σ lies on Σ (with one dimension suppressed). The dashed line is an integral curve of the vector field n^a normal to the orbit spheres



8.3 The Vacuum Schwarzschild Solution

8.3.1 Static Spherically Symmetric Metrics

Proposition 8.3.1 Suppose a static spherically symmetric spacetime (M, g_{ab}) has only one² hypersurface orthogonal timelike Killing vector field ξ^a , and G_3 is the subgroup of its isometry group that is isometric to $SO(3)$, then all of the orbit spheres of G_3 must be orthogonal to ξ^a .

Proof $\phi \in G_3$ can be viewed as an isometry from M to M . Since whether or not a vector field is timelike, Killing and hypersurface orthogonal are all determined by the metric, one can believe that $\phi_*\xi^a$ is also a hypersurface orthogonal timelike Killing vector field (see Exercise 4.12). Now that we only have one such vector field, we have $\phi_*\xi^a = \xi^a$. Assume that ξ^a is not orthogonal to an orbit sphere \mathcal{S} of G_3 , then there exists a projection $\hat{\xi}^a$ of ξ^a which is tangent to \mathcal{S} . One can always find a rotation $\hat{\phi} : \mathcal{S} \rightarrow \mathcal{S}$ on the sphere such that $\hat{\xi}^a$ will change under this rotation, i.e., $\hat{\phi}_*\hat{\xi}^a \neq \hat{\xi}^a$. However, $\hat{\phi} : \mathcal{S} \rightarrow \mathcal{S}$ can be regarded as the result of some $\phi \in G_3$ ($\phi : M \rightarrow M$) restricted to \mathcal{S} . That is, as long as $\hat{\xi}^a$ is nonvanishing, there exists a $\phi \in G_3$ such that $\phi_*\hat{\xi}^a \neq \hat{\xi}^a$, and thus $\phi_*\xi^a \neq \xi^a$, which contradicts $\phi_*\xi^a = \xi^a$. \square

Suppose Σ is a hypersurface orthogonal to ξ^a , then according to Proposition 8.3.1, an orbit surface of G_3 passing through any point of Σ lies on Σ , as shown in Fig. 8.6. Using this geometric property, we can further simplify the static line element (8.1.3). To do this we only have to specify how to define the 3-dimensional local coordinate system $\{x^1, x^2, x^3\}$ on the constant- t surface Σ . x^1 can be defined using the radius of the orbit sphere: the x^1 of each point is defined as the radius r of the orbit sphere where the point stays. x^2 and x^3 can be defined using the “carry method”: suppose \mathcal{S} is an orbit sphere in Σ , then it is a (2-dimensional) hypersurface in Σ , on which there exists a unit normal vector field n^a tangent to Σ . Since for any point on Σ there exists an orbit sphere lying on Σ that passes through the point, n^a is a vector field defined on Σ whose integral curves (one of them is shown as the dashed line in Fig. 8.6) are everywhere orthogonal to the orbit spheres. By choosing any spherical coordinates θ and φ on \mathcal{S} , we can “carry” these two coordinates to the other orbit spheres by means of the integral curves of n^a (that is, setting the values of θ and φ at

² Of course, ξ^a multiplied by an arbitrary constant is also a Killing vector field. Here by “one” we mean “one linearly independent”.

each point on each integral curve as the values of them at the intersection of \mathcal{S} and this curve), then we get a local coordinate system $\{r, \theta, \varphi\}$ on Σ . In this coordinate system, $g_{ij}dx^i dx^j$ in (8.1.3) takes the simplest form. From the above definition of θ and φ we can see that the integral curves of the normal vector field coincide with the r -coordinate lines (only with different parameters), and thus $g_{ab}(\partial/\partial r)^a (\partial/\partial \theta)^b = 0$, $g_{ab}(\partial/\partial r)^a (\partial/\partial \varphi)^b = 0$. Hence, the coefficients of the terms $dr d\theta$ and $dr d\varphi$ in $g_{ij}dx^i dx^j$ vanish. Also considering that the induced metric of $g_{ij}dx^i dx^j$ on each orbit sphere is given by (8.2.5), we have

$$g_{ij}dx^i dx^j = g_{11}dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2),$$

and therefore

$$ds^2 = g_{00}dt^2 + g_{11}dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (8.3.1)$$

According to (8.1.3), neither g_{00} nor g_{11} is a function of t . Considering the spherical symmetry, we can believe that g_{00} and g_{11} are not functions of θ or φ either [motivated readers may prove this using the property that θ and φ are constants on the integral curves of $(\partial/\partial r)^a$ and $(\partial/\partial t)^a$]. Denote g_{00} and g_{11} as $-e^{2A(r)}$ and $e^{2B(r)}$, respectively, then (8.3.1) becomes

$$ds^2 = -e^{2A(r)}dt^2 + e^{2B(r)}dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (8.3.2)$$

This is a quite general line element expression for a spherically symmetric metric that has a unique static Killing vector field in the above coordinate system $\{t, r, \theta, \varphi\}$. We emphasize that $\{t, r, \theta, \varphi\}$ is a local coordinate system in M , by which we mean that its domain (coordinate patch) cannot be the whole manifold M . Surely, even the coordinates θ and φ on each orbit sphere cannot be defined on the whole sphere (one cannot use a coordinate system to cover the whole S^2 , see Sect. 2.1). Moreover, for instance, a point where $(dr)_a = 0$ is not in the coordinate patch of $\{t, r, \theta, \varphi\}$ (the point at $X = T = 0$ in Fig. 9.13 is such a point).

8.3.2 The Vacuum Schwarzschild Solution

The static spherically symmetric metric satisfying the vacuum Einstein equation is called the **vacuum Schwarzschild solution**, or Schwarzschild solution for short, which in physics describes the outer gravitational field of a spherically symmetric star (e.g., the Sun). We have pointed out in Chap. 7 that the vacuum Einstein equation is equivalent to (see Exercise 7.7)

$$R_{ab} = 0. \quad (8.3.3)$$

Since the general form of a static spherically symmetric metric (line element) (8.3.2) only contains two undetermined functions of one variable, namely $A(r)$ and $B(r)$,

solving this equation now becomes simple: one can just express the Ricci tensor R_{ab} in terms of these two functions, set it to zero, and then solve for $A(r)$ and $B(r)$ from the resulting differential equations. In Sect. 5.7 we have introduced in detail the method and outcomes of computing the Riemann tensor of the line element (8.3.2) using the orthonormal tetrad, from which we can easily obtain the expression of R_{ab} in terms of $A(r)$ and $B(r)$. To help the readers to better understand the coordinate basis method of computing the curvature, here we compute R_{ab} again directly using the coordinate basis. First we compute the Christoffel symbols of the line element (8.3.1). It follows from (3.4.19) that the nonvanishing Christoffel symbols are

$$\begin{aligned}\Gamma^0_{01} &= \Gamma^0_{10} = A', & \Gamma^1_{00} &= A'e^{2(A-B)}, & \Gamma^1_{11} &= B', \\ \Gamma^1_{22} &= -re^{-2B}, & \Gamma^1_{33} &= -r \sin^2 \theta e^{-2B}, & \Gamma^2_{12} &= \Gamma^2_{21} = \frac{1}{r}, \\ \Gamma^2_{33} &= -\sin \theta \cos \theta, & \Gamma^3_{13} &= \Gamma^3_{31} = \frac{1}{r}, & \Gamma^3_{23} &= \Gamma^3_{32} = \cot \theta,\end{aligned}\quad (8.3.4)$$

where ' stands for the derivative with respect to r . Plugging (8.3.4) into (3.4.21) we find that the nonvanishing $R_{\mu\nu}$ are

$$R_{00} = -e^{2(A-B)}(-A'' + A'B' - A'^2 - 2r^{-1}A'), \quad (8.3.5)$$

$$R_{11} = -A'' + A'B' - A'^2 + 2r^{-1}B', \quad (8.3.6)$$

$$R_{22} = -e^{-2B}[1 + r(A' - B')] + 1, \quad (8.3.7)$$

$$R_{33} = -\{e^{-2B}[1 + r(A' - B')] - 1\} \sin^2 \theta. \quad (8.3.8)$$

Thus, $R_{ab} = 0$ is equivalent to the following three differential equations for the undetermined functions $A(r)$ and $B(r)$ [Equations (8.3.7) and (8.3.8) give the same equation]:

$$-A'' + A'B' - A'^2 - 2r^{-1}A' = 0, \quad (8.3.9)$$

$$-A'' + A'B' - A'^2 + 2r^{-1}B' = 0, \quad (8.3.10)$$

$$-e^{-2B}[1 + r(A' - B')] + 1 = 0. \quad (8.3.11)$$

Subtracting (8.3.10) from (8.3.9) yields

$$A' = -B', \quad (8.3.12)$$

and hence

$$A = -B + \alpha, \quad \alpha = \text{constant}. \quad (8.3.13)$$

Noticing (8.3.12), (8.3.11) can be rewritten as an equation with only one undetermined function $B(r)$:

$$1 - 2rB' = e^{2B}, \quad (8.3.14)$$

whose general solution is

$$e^{2B} = \left(1 + \frac{C}{r}\right)^{-1}, \quad (8.3.15)$$

where C is a constant of integration. By a direct check we can see that (8.3.13) and (8.3.15) also satisfy (8.3.9) and (8.3.10), and hence they are the general solutions of the unsolved equations (8.3.9)–(8.3.11). Plugging the A and B in these two results into the line element (8.3.2) yields

$$ds^2 = -\left(1 + \frac{C}{r}\right)e^{2\alpha} dt^2 + \left(1 + \frac{C}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (8.3.16)$$

Defining a new coordinate $\hat{t} := e^\alpha t$, we obtain

$$ds^2 = -\left(1 + \frac{C}{r}\right)d\hat{t}^2 + \left(1 + \frac{C}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (8.3.17)$$

The fact that α is a constant assures that $(\partial/\partial\hat{t})^a$ is a Killing vector field just like $(\partial/\partial t)^a$. One may choose \hat{t} to be the Killing time coordinate in the first place when the coordinate system $\{t, r, \theta, \varphi\}$ was defined, then the \hat{t} in (8.3.17) can be simply written as t :

$$ds^2 = -\left(1 + \frac{C}{r}\right)dt^2 + \left(1 + \frac{C}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (8.3.17')$$

This is the vacuum Schwarzschild solution (Schwarzschild metric). When r is sufficiently large, the equation above will approximately return to the expression for the Minkowski line element in a spherical coordinate system, and thus the Schwarzschild metric is asymptotically flat. However, when $r \rightarrow \infty$, (8.3.16) can only approach

$$ds^2 = -e^{2\alpha} dt^2 + dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2),$$

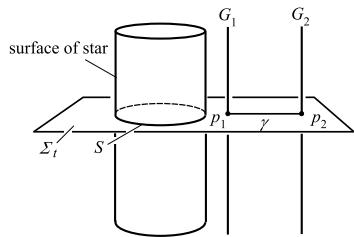
which shows one of the benefits of choosing \hat{t} as the time coordinate in the first place.

When r is sufficiently large, the linearized approximation of general relativity (see Sect. 7.8.1) can be applied. Also, $(1 + C/r)^{-1} \cong 1 - C/r$, and hence (8.3.17') approximately gives

$$ds^2 = [-dt^2 + dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2)] - \frac{C}{r}(dt^2 + dr^2).$$

The first term on the right-hand side of the above equation is a flat line element, which can be rewritten as $[-dt^2 + dx^2 + dy^2 + dz^2]$ by a coordinate transformation $x = r \sin \theta \cos \varphi$, $y = r \sin \theta \sin \varphi$, $z = r \cos \theta$. Thus, the Schwarzschild metric can be expressed as $g_{ab} = \eta_{ab} + \gamma_{ab}$ when r is large, where the 00- (i.e., tt -) component

Fig. 8.7 The spatial distance between static observers G_1 and G_2 at t is the arc length of the geodesic γ (lying on Σ_t) between p_1 and p_2



of the small quantity γ_{ab} is $\gamma_{00} = -C/r$. Comparing with (7.8.35) we get $\phi = C/2r$, and from Newton's theory of gravity we also know that $\phi = -M/r$ (where M is the mass of the star). Therefore, $C = -2M$, and hence (8.3.17') can be expressed as

$$ds^2 = -\left(1 - \frac{2M}{r}\right)dt^2 + \left(1 - \frac{2M}{r}\right)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2). \quad (8.3.18)$$

This is the most common expression of the vacuum Schwarzschild solution, in which M is the mass of the star. For a precise understanding of the concept of the “mass of a star”, see Optional Reading 9.3.1 and Chap. 12 in Volume II.

Now, let us discuss the Schwarzschild metric in a more “physical” manner, i.e., we will discuss the spatial geometry outside a static spherically symmetric star. The cylindrical surface in Fig. 8.7 represents the world sheet of the surface of a static spherically symmetric star, and the spacetime geometry outside this surface is described by the Schwarzschild metric. There exists a static reference frame in Schwarzschild spacetime, in which each constant- t surface Σ_t can be interpreted as the space in this reference frame at t . The intersecting surface S of Σ_t and the cylindrical surface represents the surface of the star at t (which is suppressed as a 1-dimensional circle in the figure). Suppose G_1 and G_2 are two static observers who have the same values of θ and φ , and the intersections p_1 and p_2 of their world lines and Σ_t represent the positions of these two observers at t . The spatial geometry outside of S in Σ is described by the induced metric h_{ab} of the Schwarzschild metric; the corresponding line element is

$$ds^2 = \left(1 - \frac{2M}{r}\right)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2). \quad (8.3.19)$$

Let us compute the spatial distance l between p_1 and p_2 . The distance between two points in a Riemannian space (the metric is positive definite) is defined as the arc length of the shortest curve among all the curves connecting these two points.³ It is not difficult to show that the curve γ on Σ_t from p_1 to p_2 with θ and φ being constants is the shortest curve between p_1 and p_2 , whose length (and thus the distance between p_1 and p_2) is

³ Technically speaking, the **distance** between two points in a Riemannian space is defined as the infimum of the set of the lengths of all the curves between these two points (as a subset of \mathbb{R}).

$$l = \int (h_{ij} dx^i dx^j)^{1/2} = \int_{r_1}^{r_2} \left(1 - \frac{2M}{r}\right)^{-1/2} dr > r_2 - r_1,$$

where r_1 and r_2 are the r -coordinates of G_1 and G_2 , respectively. The equation above indicates that the spatial distance between G_1 and G_2 at any time t is a constant (which is a property of static observers). l is also called the **proper distance** between G_1 and G_2 , which is not equal to their coordinate distance $r_2 - r_1$. This is exactly a reflection of (Σ_t, h_{ab}) being non-Euclidean.

In this chapter, the main point regarding the Schwarzschild metric is about finding the solution from Einstein's equation. We will have a detailed discussion on Schwarzschild spacetime later in Chap. 9.

To facilitate future lookup, here we list the components of the Christoffel symbol and the Riemann tensor (with lower indices) of the Schwarzschild metric in the Schwarzschild coordinate system as follows (in which x^0, x^1, x^2, x^3 stand for t, r, θ, φ , respectively):

$$\left. \begin{aligned} \Gamma^0_{01} = \Gamma^0_{10} &= \frac{M}{r^2}(1-2M/r)^{-1}, & \Gamma^1_{00} &= \frac{M}{r^2}(1-2M/r), \\ \Gamma^1_{11} &= -\frac{M}{r^2}(1-2M/r)^{-1}, & \Gamma^1_{22} &= -r(1-2M/r), & \Gamma^1_{33} &= -r(1-2M/r)\sin^2\theta, \\ \Gamma^2_{12} = \Gamma^2_{21} &= \frac{1}{r}, & \Gamma^2_{33} &= -\sin\theta\cos\theta, & \Gamma^3_{13} = \Gamma^3_{31} &= \frac{1}{r}, & \Gamma^3_{23} = \Gamma^3_{32} &= \cot\theta, \end{aligned} \right\} \quad (8.3.20)$$

$$\left. \begin{aligned} R_{0101} &= -\frac{2M}{r^3}, & R_{0202} &= \frac{M}{r}(1-2M/r), & R_{0303} &= \frac{M}{r}(1-2M/r)\sin^2\theta, \\ R_{1212} &= -\frac{M}{r}(1-2M/r)^{-1}, & R_{1313} &= -\frac{M}{r}(1-2M/r)^{-1}\sin^2\theta, & R_{2323} &= 2Mr\sin^2\theta. \end{aligned} \right\} \quad (8.3.21)$$

[Optional Reading 8.3.1]

We have repeatedly mentioned that “gravity is an effect of curved spacetime”. Now that we have introduced the concept of a stationary spacetime, we can provide a deeper and more specific interpretation for this. The earliest concept of gravity came from the study of the motion of objects near the Earth. When you release an apple in your hand, it will fall to the ground with an acceleration $|\vec{g}| = 9.8 \text{ m} \cdot \text{s}^{-2}$, and so we say that the apple experiences the Earth's gravity, or the Earth produces a gravitational field outside itself. What does such an important quantity $|\vec{g}|$ correspond to in general relativity? From the perspective of general relativity, the apple undergoes geodesic motion with vanishing 4-acceleration. In contrast, although you (as a stationary observer) feel that you are sitting comfortably in a chair, your 4-acceleration is nonzero. The 4-acceleration of a stationary observer is (see Exercise 8.3)

$$A^a = \nabla^a \ln \chi, \quad (8.3.22)$$

where $\chi \equiv (-\xi_a \xi^a)^{1/2}$, and ξ^a is the timelike Killing vector field of the stationary spacetime. Since the 4-acceleration is orthogonal to the 4-velocity, A^a is a spatial vector field on the world line of the stationary observer. This is an intrinsic vector field of the stationary spacetime geometry itself. The gravitational field strength \vec{g} in the Newtonian language must correspond to a certain intrinsic geometric quantity in general relativity. $-A^a$ is exactly

such a quantity, and thus can be called the “gravitational field” (gravitational acceleration field) in the stationary spacetime. Now we will show that this terminology indeed agrees with the value of the gravitational field strength $|\vec{g}| = 9.8 \text{ m} \cdot \text{s}^{-2}$ in your mind. Consider approximately that there is a Schwarzschild metric outside the Earth, then

$$\chi \equiv (-\xi_a \xi^a)^{1/2} = (-g_{00})^{1/2} = (1 - 2M/r)^{1/2},$$

and (8.3.22) becomes

$$A_a = \chi^{-1} \nabla_a \chi = \frac{M}{r^2} (1 - 2M/r)^{-1} (\mathrm{d}r)_a.$$

Thus,

$$|A^a| = \sqrt{g_{ab} A^a A^b} = \frac{M}{r^2} (1 - 2M/r)^{-1} \sqrt{g^{ab} (\mathrm{d}r)_a (\mathrm{d}r)_b} = \frac{M}{r^2} (1 - 2M/r)^{-1} \sqrt{g^{11}},$$

and hence

$$|A^a| = \frac{M}{r^2} (1 - 2M/r)^{-1/2}. \quad (8.3.23)$$

Suppose the world lines of an apple G and a stationary observer G_s are tangent at p (see Fig. 8.8). Due to its free fall, G corresponds to an inertial observer in Minkowski spacetime, and it follows from Proposition 6.3.6 and the equivalence principle that the 3-acceleration a^a of G at p relative to G_s is equal to the negative of the (absolute) 4-acceleration of G_s . Also since a^a is \vec{g} , changing (8.3.23) back to the International System of Units (SI) we have

$$|\vec{g}| = |A^a| = \frac{GM}{r^2} \left(1 - \frac{2GM}{c^2 r}\right)^{-1/2}. \quad (8.3.24)$$

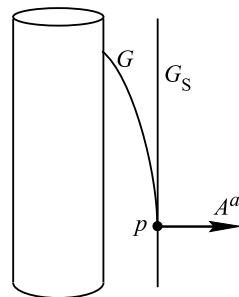
Applying this to the Earth’s surface, and plugging in $M = M_\oplus = 6 \times 10^{24}$, $r = r_\oplus = 6.4 \times 10^6$, $c = 3 \times 10^8$, $G = 6.7 \times 10^{-11}$, we find that the parentheses on the right-hand side of the above equation equals $1 - 10^{-9} \cong 1$. Hence,

$$|\vec{g}| \cong \frac{GM_\oplus}{r_\oplus^2} \cong 9.8.$$

Therefore, we say that there exists a gravitational field $-A^a = -\nabla^a \ln \chi$ in a stationary spacetime, which is the general relativity formulation for the Earth’s gravitational field \vec{g} . However, a new question arises: there is no stationary observer in a non-stationary spacetime, and gravity in the above sense does not exist, so how do we interpret the statement “a curved spacetime must have gravity”? As we have mentioned, as long as the spacetime is curved, there will be a geodesic deviation effect (tidal effect), which can be referred to as a relative gravitational effect. This effect is inherent to curved spacetime, which is different from the gravitational effect in the former sense. (That is, in a stationary spacetime one can always eliminate gravity in the first sense by choosing a freely falling elevator, but one cannot eliminate the tidal effect). In fact, the geodesic deviation effect is a common property that all the curved spacetimes share. When saying “a curved spacetime must have gravity”, for a non-stationary spacetime this is referring to the relative gravity (tidal effect) between freely falling bodies. When the spacetime curvature is everywhere vanishing, gravity in either sense does not exist, and therefore we say “there is no gravity without curved spacetime”.

[The End of Optional Reading 8.3.1]

Fig. 8.8 A freely falling apple G has a 3-acceleration $a^a = -A^a$ relative to the stationary observer G_s



[Optional Reading 8.3.2]

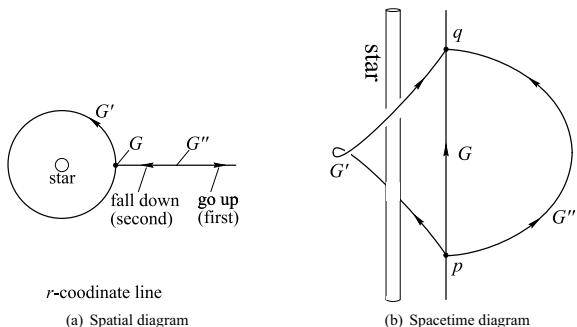
Suppose $\{t, r, \theta, \varphi\}$ is a Schwarzschild coordinate system outside an isolated static spherically symmetric star, and G is a static observer outside the star, whose spatial coordinate values are $r = r_G$, $\theta = \pi/2$ and $\varphi = 0$; G' is a free observer undergoing circular motion around the star due to the star's gravitational field, whose θ value is always $\pi/2$ (always right above the star's equator). At the beginning ($\tau = 0$), the world lines of G' and G intersect at p , and they intersect again at q after G' goes around the star (see Fig. 8.9). It is easy to see from (8.3.18) that the contribution from dt to the line elements of G and G' are equal, while the line element of G' has another contribution from $d\varphi$, namely $r^2 \sin^2 \theta d\varphi^2$, which has an opposite sign. Hence, the curve G' between p and q is shorter than G . How could a timelike geodesic G' be shorter than a non-geodesic G ? First of all, “the length of a timelike geodesic is a maximum” is talking about the comparison among the infinitesimally nearby timelike curves (“local maximum”), while G and G' are not nearby. Secondly, there does exist a timelike curve that is infinitesimally close to G' and longer than G' , which is not surprising since the necessary and sufficient condition for “the length of a timelike geodesic is a maximum” to hold is that there does not exist a pair of conjugate points on the curve, and G' does not satisfy this condition. In fact, we can believe that (one can prove this based on the definition of conjugate points in Optional Reading 7.6.3) there exist infinitely many pairs of conjugate points on G' [two points satisfying $\varphi = \varphi_0$ and $\varphi = \varphi_0 + \pi$ (where $0 \leq \varphi_0 < \pi$) make a pair]. A timelike geodesic between p and q without conjugate points corresponds to the following physical situation: suppose a free observer G'' is projected straight up with some initial speed at an event p and falls freely (follows a radial geodesic), and then meets the curve G again at q (see Fig. 8.9), then its world line is at least a local maximum due to the nonexistence of conjugate points. It is actually longer than both G and G' . [Note that from (8.3.18) one can tell that the argument for G' being shorter than G does not apply to G'' since its r -value does not equal the r -value of G and so the contribution from dt to these line elements are not equal].

[The End of Optional Reading 8.3.2]

8.3.3 Birkhoff's Theorem

Schwarzschild showed that the static spherically symmetric solution to the vacuum Einstein equation is the Schwarzschild solution, as we have introduced above. Later it was found that the static condition can actually be removed, because in 1923 G. D. Birkhoff proved the following theorem: a spherically symmetric solution to the vacuum Einstein equation must be static. Here we briefly sketch the idea of the

Fig. 8.9 Observers G , G' and G'' part from each other at an event p and reunite at q . G' and G'' are geodesics, and their arc lengths have the relation $l_{G'} < l_G < l_{G''}$



proof. The general form of a static spherically symmetric line element is (8.3.2). If one removes the static condition, the expression for the line element will not be as simple, for example the coefficient of the cross term $d\tau/dr$ will be nonzero. However, by an appropriate coordinate transformation, one can change the line element to the same form as (8.3.2), and the only difference is that the functions of one variable $A(r)$ and $B(r)$ now become functions of two variables $A(t, r)$ and $B(t, r)$. Let A' , B' , \dot{A} and \dot{B} represent $\partial A/\partial r$, $\partial B/\partial r$, $\partial A/\partial t$ and $\partial B/\partial t$, respectively. Through a procedure which is slightly more complicated than the computation in Sect. 8.3.2 [see Carmeli (1982); Stephani (1982)], we will still obtain the Schwarzschild line element (8.3.18).

Birkhoff's theorem is a powerful theorem, which asserts that as long as a non-static matter distribution keeps being spherical symmetric (such as a star that is sharply contracting, expanding, oscillating, or even exploding in the radial direction), the external spacetime geometry will still be described by the vacuum Schwarzschild solution. This provides great convenience for the study of stellar evolution (see Sects. 9.3 and 9.4).

Birkhoff's theorem is very similar to the following theorem in electrodynamics: the electromagnetic field of a spherically symmetric charge distribution (i.e., a spherically symmetric solution to the vacuum Maxwell equations) must be an electrostatic field. An electromagnetic wave is the propagation of a time-dependent electromagnetic field in space, and “a spherically symmetric electromagnetic field must be an electrostatic field” indicates that there does not exist any spherically symmetric electromagnetic wave. (A spherical electromagnetic wave is an electromagnetic wave whose wavefront is a sphere; its electromagnetic field does not have spherical symmetry, and thus it is not a spherically symmetric electromagnetic wave). Similarly, since a gravitational wave will not appear in a stationary gravitational field (stationary means time-independent), Birkhoff's theorem indicates that there does not exist any spherically symmetric gravitational wave. Noticing that spherically symmetric radiation is monopole radiation, an equivalent statement of the conclusion above is: there does not exist monopole electromagnetic or gravitational radiation. The major contribution of electromagnetic radiation comes from dipole radiation. In contrast, from Sect. 7.9 we can see that for gravity there exists neither monopole radiation

nor dipole radiation. The major contribution of gravitational radiation comes from quadrupole radiation. Table 8.1 provides a comparison between these two kinds of radiation.

Later, it was found that the original formulation by Birkhoff was not precise enough. The revised Birkhoff's theorem can be formulated as follows: a spherically symmetric solution to the vacuum Einstein equation must be the Schwarzschild metric. The difference between this revised version and the original version is that the extended Schwarzschild metric will be non-stationary in some spacetime region, see Sect. 9.4.3 for details. The original Birkhoff's theorem was first challenged by A. Z. Petrov in 1963 [see Stephani et al. (2003) p. 232 and the references therein]. For a proof of the revised Birkhoff's theorem, see Appendix B of Hawking and Ellis (1973). Kuang and Liang (1988) further generalized this theorem by weakening the spherical symmetry condition to “conformally spherical symmetry”. The definition of the term “conformal” will be introduced in Sect. 12.1 (Volume II).

8.4 The Reissner-Nordström Solution

8.4.1 Electrovacuum Spacetimes and the Einstein-Maxwell Equations

The Schwarzschild metric describes the curved spacetime (vacuum) outside a static spherically symmetric star. Many actual stars (or celestial bodies) carry electric charges, and their exterior spacetime is not vacuum but filled with an electromagnetic field. A spacetime with only an electromagnetic field but without a matter field is called an **electrovacuum** (or electrovac for short) **spacetime**. The T_{ab} in the electrovacuum Einstein equation $G_{ab} = 8\pi T_{ab}$ is the energy-momentum tensor for some electromagnetic field F_{ab} (we will only talk about source-free electromagnetic fields), i.e.,

$$T_{ab} = \frac{1}{4\pi} (F_{ac} F_b^c - \frac{1}{4} g_{ab} F_{cd} F^{cd}). \quad (8.4.1)$$

Hence, the electrovacuum Einstein equation can also be expressed as

$$G_{ab} \equiv R_{ab} - \frac{1}{2} R g_{ab} = 2(F_{ac} F_b^c - \frac{1}{4} g_{ab} F_{cd} F^{cd}), \quad (8.4.2)$$

Table 8.1 Comparative table for gravitational radiation and electromagnetic radiation

	Monopole radiation	Dipole radiation	Quadrupole radiation
Electromagnetic radiation	Nonexistent	Exists (major)	Exists
Gravitational radiation	Nonexistent	Nonexistent	Exists (major)

where F_{ab} satisfies the source-free Maxwell equations in curved spacetime

$$\nabla^a F_{ab} = 0, \quad (8.4.3a)$$

$$\nabla_{[a} F_{bc]} = 0. \quad (8.4.3b)$$

Here ∇_a is the derivative operator associated with the metric g_{ab} , and g_{ab} must satisfy (8.4.2). Thus, an electrovacuum spacetime is determined by three ingredients: a background manifold M , a metric field g_{ab} and an electromagnetic field F_{ab} , among which g_{ab} and F_{ab} are the solutions of the simultaneous equations formed by (8.4.2) and (8.4.3). This system of equations is called the **Einstein-Maxwell equations**. It is easy to show from (8.4.1) that (Exercise 8.4) the trace of the energy-momentum tensor T_{ab} of the electromagnetic field is $T \equiv g^{ab} T_{ab} = 0$, and hence from Einstein's equation $R_{ab} - \frac{1}{2} R g_{ab} = 8\pi T_{ab}$ one can easily see that (Exercise 8.4) the scalar curvature $R = 0$. Therefore, the electrovacuum Einstein equation can be simplified as

$$R_{ab} = 8\pi T_{ab}. \quad (8.4.4)$$

Based on their physical properties, electromagnetic fields F_{ab} can be classified into null electromagnetic fields and nonnull electromagnetic fields. Define a complex tensor field

$$\Sigma_{ab} := F_{ab} + i^* F_{ab}, \quad (8.4.5)$$

where ${}^* F_{ab}$ is the Hodge dual of F_{ab} . F_{ab} is called a **null electromagnetic field** if

$$\Sigma_{ab} \Sigma^{ab} = 0, \quad (8.4.6)$$

otherwise it is called a **nonnull electromagnetic field**. It is easy to show that (Exercise 8.5)

$$\Sigma_{ab} \Sigma^{ab} = 2(F_{ab} F^{ab} + i F_{ab} {}^* F^{ab}), \quad (8.4.7)$$

and thus the null condition (8.4.6) of an electromagnetic field is equivalent to

$$F_{ab} F^{ab} = 0, \quad (8.4.8a)$$

and

$$F_{ab} {}^* F^{ab} = 0. \quad (8.4.8b)$$

The electric field and magnetic field measured by an instantaneous observer (p, Z^a) at a point p are by definition $E := F_{ab} Z^b$ and $B_a := -{}^* F_{ab} Z^b$ (see Sect. 6.1.1), from which one can show that (see Exercise 6.15)

$$F_{ab} F^{ab} = 2(B^2 - E^2), \quad (8.4.9)$$

$$F_{ab} {}^* F^{ab} = 4\vec{E} \cdot \vec{B} \equiv 4g^{ab} E_a B_b. \quad (8.4.10)$$

Thus, although both \vec{E} and \vec{B} depend on the observer, $B^2 - E^2$ and $\vec{E} \cdot \vec{B}$ are two invariants (i.e., scalar fields). (In fact, these are the only two independent invariants that one can construct out of \vec{E} and \vec{B}). The two equations above indicate that (8.4.8) is equivalent to

$$B^2 = E^2, \quad (8.4.11a)$$

$$\vec{E} \cdot \vec{B} = 0. \quad (8.4.11b)$$

These two equations indicate that the \vec{E} and \vec{B} measured by an instantaneous observer are orthogonal and have the same magnitude, which are exactly the two basic properties of an electromagnetic plane wave in Minkowski spacetime. It can be proved that (see Appendix D in Volume II), suppose in an arbitrary spacetime there exists a null electromagnetic field F_{ab} whose energy-momentum tensor is T_{ab} , then the 4-momentum density $W^a \equiv -T^a{}_b Z^b$ of F_{ab} (see Sect. 6.4) measured by an instantaneous observer (p, Z^a) is a future-directed null vector.

8.4.2 The Reissner-Nordström Solution

Now we will solve the Einstein-Maxwell equations of a static spherically symmetric star. According to the discussion in Sect. 8.3.1, in the static spherically symmetric case one can choose a coordinate system $\{x^\mu\} \equiv \{t, r, \theta, \varphi\}$ adapted to two geometric properties (staticity and spherical symmetry) of the metric and express the line element as the following simple form [i.e., (8.3.2)]:

$$ds^2 = -e^{2\alpha(r)} dt^2 + e^{2\beta(r)} dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (8.4.12)$$

[The $A(r)$ and $B(r)$ in (8.3.2) may be confused with the 4-potential A and the magnetic field B , and hence we now denote them by $\alpha(r)$ and $\beta(r)$.] This coordinate system can not only simplify the line element, but also simplify the components of the electromagnetic field. The electromagnetic field F_{ab} produced by a charged static spherically symmetric star is also static and spherically symmetric. The components A_μ of its electromagnetic 4-potential A_a are independent of the coordinates t, θ, φ , and there is no component tangent to the orbit sphere, i.e., $A_2 = A_3 = 0$. Note that A_a has a gauge freedom: suppose χ is an arbitrary function of r , then $\tilde{A}_a = A_a + \nabla_a \chi$ and A_a correspond to the same F_{ab} . From this equation we get

$$\tilde{A}_1 = (\partial/\partial r)^a (A_a + \nabla_a \chi) = A_1 + \partial \chi / \partial r.$$

Thus, for any given A_a one can always choose a suitable $\chi(r)$ such that $\tilde{A}_1 = 0$, and hence A_0 can be regarded as the only component of A_a . Also from

$$F_{\mu\nu} = 2\partial_{[\mu} A_{\nu]} = \partial_\mu A_\nu - \partial_\nu A_\mu$$

we can see that the only nonvanishing $F_{\mu\nu}$ are

$$-F_{01} = F_{10} = \partial_1 A_0 = \frac{dA_0}{dr}, \quad (8.4.13)$$

i.e., F_{ab} only has one independent component F_{01} , whose expression can be obtained by solving Maxwell's equations (8.4.3). Equation (8.4.3b) is automatically satisfied since it follows from $\mathbf{F} = d\mathbf{A}$ that $d\mathbf{F} = d(d\mathbf{A}) = 0$. The coordinate component form of (8.4.3a) reads

$$F^{\mu\nu}_{;\mu} = 0, \quad \nu = 0, 1, 2, 3. \quad (8.4.14)$$

Using a similar way of deriving (3.4.26) we get

$$F^{\mu\nu}_{;\mu} = \frac{1}{\sqrt{-g}} \frac{\partial}{\partial x^\mu} (\sqrt{-g} F^{\mu\nu}) + \Gamma^\nu_{\sigma\mu} F^{\mu\sigma} = \frac{1}{\sqrt{-g}} \frac{\partial}{\partial x^\mu} (\sqrt{-g} F^{\mu\nu}), \quad (8.4.15)$$

and it follows from (8.4.13) and (8.4.12) that the only nonvanishing $\sqrt{-g} F^{\mu\nu}$ are $\sqrt{-g} F^{01} = -\sqrt{-g} F^{10} = r^2 F_{10} e^{-(\alpha+\beta)} \sin \theta$. Hence, when $\nu = 1, 2, 3$, (8.4.14) are identities, and when $\nu = 0$ it gives

$$\frac{d}{dr} [r^2 F_{10}(r) e^{-\alpha(r)-\beta(r)}] = 0,$$

whose general solution is

$$F_{10} = \frac{Q}{r^2} e^{\alpha+\beta}, \quad \text{where } Q = \text{constant}. \quad (8.4.16)$$

So far, an electromagnetic field F_{ab} satisfying Maxwell's equations has the following expression:

$$F_{ab} = -\frac{Q}{r^2} e^{\alpha+\beta} (dt)_a \wedge (dr)_b. \quad (8.4.17)$$

The equation above still contains undetermined functions $\alpha(r)$ and $\beta(r)$, which should be obtained from Einstein's equation (8.4.4). From the very beginning, we have two sets of undetermined functions, namely $\{F_{\mu\nu}(r)\}$ and $\{\alpha(r), \beta(r)\}$. Do not naively think that the former only appears in Maxwell's equations and the latter only appears in Einstein's equation, so that they can be solved independently. In truth, both of them appear in both sets of equations, and thus the Einstein-Maxwell equations are coupled equations, which means they are interdependent on each other. Now we will solve Einstein's equation $R_{ab} = 8\pi T_{ab}$. In order to do so, first we compute the energy-momentum tensor T_{ab} of F_{ab} . It follows from (8.4.1) and (8.4.12) that the nonvanishing coordinate components of T_{ab} are

$$\begin{aligned} T_{00} &= F_{10}^2 e^{-2\beta} / 8\pi, & T_{11} &= -F_{10}^2 e^{-2\alpha} / 8\pi, \\ T_{22} &= r^2 F_{10}^2 e^{-2(\alpha+\beta)} / 8\pi, & T_{33} &= r^2 F_{10}^2 e^{-2(\alpha+\beta)} \sin^2 \theta / 8\pi. \end{aligned} \quad (8.4.18)$$

On the other hand, the expressions for the nonvanishing coordinate components $R_{\mu\nu}$ of the Ricci tensor R_{ab} are given by (8.3.5)–(8.3.8), and hence the component equations for Einstein's equation (8.4.4), $R_{00} = 8\pi T_{00}$ and $R_{11} = 8\pi T_{11}$, are equivalent to

$$-e^{2(\alpha-\beta)}(-\alpha'' + \alpha'\beta' - \alpha'^2 - 2r^{-1}\alpha') = F_{10}^2 e^{-2\beta}, \quad (8.4.19)$$

$$-\alpha'' + \alpha'\beta' - \alpha'^2 + 2r^{-1}\beta' = -F_{10}^2 e^{-2\alpha}. \quad (8.4.20)$$

We can easily get from the two equations above that $\alpha' = -\beta'$, which is the same as (8.3.12) in the process of finding the Schwarzschild solution; hence, here we can also set $\alpha = -\beta$ by redefining t . Under this premise, we can see from (8.4.16) that the remaining two component equations $R_{22} = 8\pi T_{22}$ and $R_{33} = 8\pi T_{33}$ are equivalent to

$$(re^{2\alpha})' = 1 - \frac{Q^2}{r^2}.$$

Hence,

$$e^{2\alpha} = 1 + \frac{Q^2}{r^2} + \frac{C}{r}, \quad (8.4.21)$$

and thus

$$e^{2\beta} = \left(1 + \frac{Q^2}{r^2} + \frac{C}{r}\right)^{-1}. \quad (8.4.22)$$

Plugging into (8.4.12) yields the spacetime line element

$$ds^2 = -\left(1 + \frac{Q^2}{r^2} + \frac{C}{r}\right)dt^2 + \left(1 + \frac{Q^2}{r^2} + \frac{C}{r}\right)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \quad (8.4.23)$$

and plugging $\alpha = -\beta$ into (8.4.16) yields

$$F_{10} = \frac{Q}{r^2}. \quad (8.4.24)$$

One can now check that these expressions for α , β and F_{10} do satisfy (8.4.19) and (8.4.20). When r is sufficiently large, $Q^2/r^2 \ll C/r$, and hence (8.4.23) becomes approximately

$$ds^2 \cong -\left(1 + \frac{C}{r}\right)dt^2 + \left(1 + \frac{C}{r}\right)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2). \quad (8.4.25)$$

From the physical perspective, when r is sufficiently large, the gravitational field of a charged spherically symmetric star should approximately obey Newton's theory of gravity, and the spacetime metric should be approximately the same as the Schwarzschild metric, and thus $C = -2M$. On the other hand, the star can be viewed

as a point charge when r is sufficiently large, and the F_{10} it produces should be equal to its electric charge divided by r^2 , and hence from (8.4.24) we can see that the physical meaning of the constant Q is the electric charge of the star. Therefore, ultimately (8.4.23) can be written as

$$ds^2 = -\left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right)dt^2 + \left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \quad (8.4.26)$$

which is called the **Reissner-Nordström line element** (or **RN line element** for short). It describes the exterior spacetime geometry of a static spherically symmetric star (object) with a mass M and electric charge Q , whose corresponding electromagnetic field F_{ab} and 4-potential A_a are

$$F_{ab} = -\frac{Q}{r^2}(dt)_a \wedge (dr)_a, \quad A_a = -\frac{Q}{r}(dt)_a. \quad (8.4.27)$$

The metric g_{ab} expressed by (8.4.26) together with the electromagnetic field expressed by (8.4.27) form the RN solution of the Einstein-Maxwell equations.

Now let us have some discussion on the electromagnetic field of the RN solution. From (8.4.27) we can easily obtain $F_{ab}F^{ab} = -2Q^2/r^4 \neq 0$, and thus the F_{ab} of RN spacetime is a nonnull electromagnetic field. People always say that the electromagnetic field of RN spacetime is an electrostatic field. To understand this statement, one should notice that an observer needs to be specified when talking about an electric field and magnetic field. Now we will show that the electric field and magnetic field for the F_{ab} of an RN solution measured by a *static* observer G are, respectively, an electrostatic field and zero. The 4-velocity of G is

$$Z^a = f^{1/2}(\partial/\partial t)^a \quad [\text{where } f \equiv 1 - (2M/r) + Q^2/r^2].$$

Normalizing the dual coordinate basis vectors $(dr)_a, (d\theta)_a, (d\varphi)_a$, we have the orthonormal spatial triad of G :

$$(e^1)_a = f^{-1/2}(dr)_a, \quad (e^2)_a = r(d\theta)_a, \quad (e^3)_a = r \sin \theta (d\varphi)_a.$$

It is easy to show that (Exercise 8.5) the electric field $E_a \equiv F_{ab}Z^b$ and magnetic field $B_a \equiv -{}^*F_{ab}Z^b$ measured by G are $E_a = \frac{Q}{r^2}(e^1)_a$ and $B_a = 0$, or

$$E^a = \frac{Q}{r^2}(e_1)^a, \quad B^a = 0 \quad [\text{where } (e_1)^a \equiv f^{1/2}(\partial/\partial r)^a]. \quad (8.4.28)$$

Thus, the result of F_{ab} measured by a static observer in RN spacetime is an electrostatic field generated by a point charge Q and with no magnetic field, which also confirms the fact that F_{ab} is nonnull.⁴

⁴ In Volume II we will introduce the electromagnetic duality transformation, which only changes the formulation but does not change the essence of the physics. For instance, one can either say that

If we do not assume that the metric is static, i.e., we change the $\alpha(r)$ and $\beta(r)$ in (8.4.12) to $\alpha(t, r)$ and $\beta(t, r)$, then we will arrive at exactly the same result as we obtained above. [For details of the derivation, see Carmeli (1982)]. This can be regarded as a generalization of Birkhoff's theorem: the electrovacuum spherically symmetric solution to Einstein's equation must be the RN solution.

8.5 Axisymmetric Metrics [Optional Reading]

Many celestial bodies also have rotation. Due to the rotation, the symmetry of a spherically symmetric star will be degraded to axial symmetry. Moreover, an axisymmetric matter distribution will have axial symmetry whether or not it has any rotation with respect to the axis. Mathematically speaking, a metric g_{ab} is said to be **axisymmetric** if there exists a one-parameter group of isometries whose orbits (except for the fixed points) are closed spacelike curves. Thus, in an axisymmetric spacetime there exists a spatial Killing vector field ψ^a whose integral curves are closed curves. An axisymmetric metric g_{ab} is said to be stationary axisymmetric if it has a timelike Killing field ξ^a , and ξ^a commutes with the Killing field ψ^a which represents the axial symmetry:

$$[\xi, \psi]^a = 0. \quad (8.5.1)$$

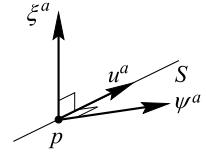
Using this commutativity we can choose a coordinate system $\{x^0 \equiv t, x^1 \equiv \varphi, x^2, x^3\}$ such that $\xi^a = (\partial/\partial t)^a$, $\psi^a = (\partial/\partial \varphi)^a$. Suppose $g_{\mu\nu}$ are the components of g_{ab} in this system, then it follows from (4.2.3) and the fact that ξ^a and ψ^a are Killing that

$$\frac{\partial g_{\mu\nu}}{\partial t} = (\mathcal{L}_\xi g)_{\mu\nu} = 0, \quad \frac{\partial g_{\mu\nu}}{\partial \varphi} = (\mathcal{L}_\psi g)_{\mu\nu} = 0, \quad (8.5.2)$$

and hence $g_{\mu\nu}$ can only be functions of x^2 and x^3 . In order to further simplify the solving process, here we only discuss the stationary axisymmetric metrics satisfying the following condition: $\forall p \in M, \exists$ a 2-dimensional surface S passing through p and orthogonal to both $\xi^a|_p$ and $\psi^a|_p$. That is, for any vector u^a at p that is tangent to S we have $g_{ab}u^a\xi^b|_p = g_{ab}u^a\psi^b|_p = 0$. (Note that since a 2-dimensional surface in a 4-dimensional spacetime is not a hypersurface, it has more than one linearly independent normal vector, see Fig. 8.10). Many important stationary axisymmetric metrics satisfy this condition. Choose an arbitrary coordinate system $\{x^2, x^3\}$ on an orthogonal surface S_0 , carry x^2 and x^3 to any point outside S_0 using the integral curves of ξ^a and ψ^a (i.e., set the x^2 and x^3 on each integral curve as constants), and set the zeros of the Killing parameters t and φ such that t and φ are constants on each orthogonal surface S (from a proposition similar to Proposition 8.1.1 we can see that this is always possible). In this way we obtain a local coordinate system $\{x^0 \equiv t, x^1 \equiv \varphi, x^2, x^3\}$, where the coordinate lines of x^0 and x^1 are the integral curves of ξ^a and ψ^a , respectively, while the coordinate lines of x^2 and x^3 lie on the orthogonal surface S . Thus, the components $g_{\mu\nu}$ of g_{ab} in this system satisfy

a charged static star carries electric charge but no magnetic charge, or say that it carries magnetic charge but no electric charge, or even say that it has both electric and magnetic charges (and the amount is flexible, as long as the sum of the squares of them is invariant). When we discuss the RN solution in this section we adopt the most common formulation, i.e., the star carries only electric charge but no magnetic charge, and the corresponding electromagnetic field has only an electrostatic field but no magnetic field.

Fig. 8.10 S is a 2-dimensional surface orthogonal to both ξ^a and ψ^a (with one dimension suppressed in the figure)



$$\begin{aligned} g_{02} = g_{20} &= g_{ab}\xi^a(\partial/\partial x^2)^b = 0, & g_{03} = g_{30} &= g_{ab}\xi^a(\partial/\partial x^3)^b = 0, \\ g_{12} = g_{21} &= g_{ab}\psi^a(\partial/\partial x^2)^b = 0, & g_{13} = g_{31} &= g_{ab}\psi^a(\partial/\partial x^3)^b = 0. \end{aligned}$$

Let $V \equiv -g_{00} = -g_{ab}\xi^a\xi^b$, $W \equiv g_{01} = g_{ab}\xi^a\psi^b$, $X \equiv g_{11} = g_{ab}\psi^a\psi^b$, then the line element can be expressed as

$$ds^2 = -Vdt^2 + Xd\varphi^2 + 2Wdtd\varphi + g_{22}(dx^2)^2 + g_{33}(dx^3)^2 + 2g_{23}dx^2dx^3. \quad (8.5.3)$$

From (8.5.2) we know that V , X , W , g_{22} , g_{33} , g_{23} can only be functions of x^2 and x^3 , and thus solving Einstein's equation can be boiled down to the problem of finding these 6 functions of two variables. However, the problem can be further simplified. Define a function ρ using the following equation:

$$\rho^2 := VX + W^2. \quad (8.5.4)$$

V , X , W are not functions of t and φ , which leads to $\xi^a\nabla_a\rho = \partial\rho/\partial t = 0$ and $\psi^a\nabla_a\rho = \partial\rho/\partial\varphi = 0$, i.e., $\nabla^a\rho$ is orthogonal to ξ^a and ψ^a , and thus is tangent to each S . We will do two things on the surface S_0 : ① choose ρ as the second coordinate x^2 , ② take any constant- ρ line and define arbitrarily a 1-dimensional coordinate z on the line, and then carry z to the other points on S_0 using the integral curves of $\nabla^a\rho$. The coordinate basis vector $(\partial/\partial\rho)^a$ of the 2-dimensional coordinate system $\{x^2 \equiv \rho, x^3 \equiv z\}$ ⁵ obtained in this way is orthogonal to $(\partial/\partial z)^a$, and hence $g_{23}|_{S_0} = 0$. Carry the x^2 and x^3 outside S_0 using the integral curves of ξ^a and ψ^a as we mentioned above, then we get a coordinate system $\{x^\mu\}$, in which $x^0 \equiv t$, $x^1 \equiv \varphi$, $x^2 \equiv \rho$, $x^3 \equiv z$. Two points needs to be elucidated: ① ρ is defined by (8.5.4), while $x^2 \equiv \rho$ is only defined on S_0 , and then we carry it outside the surface. Why do we also have $x^2 \equiv \rho$ outside the surface? This is the outcome of $\xi^a\nabla_a\rho = 0$, $\xi^a\nabla_a x^2 = 0$ (requirements of the carry method) [and the corresponding $\psi^a\nabla_a\rho = 0$, $\psi^a\nabla_a x^2 = 0$] together with $(x^2 - \rho)|_{S_0} = 0$. ② From $g_{23}|_{S_0} = 0$, $\xi^c\nabla_c g_{23} = 0$ and $\psi^c\nabla_c g_{23} = 0$ one can easily see that $g_{23} = 0$ holds on the whole coordinate patch. The proof of the latter two equations are as follows (here we only take $\xi^c\nabla_c g_{23} = 0$ as an example):

$$\begin{aligned} \xi^c\nabla_c g_{23} &= \xi^c\nabla_c[gab(\partial/\partial x^2)^a(\partial/\partial x^3)^b] = \mathcal{L}_\xi[gab(\partial/\partial x^2)^a(\partial/\partial x^3)^b] \\ &= gab[\mathcal{L}_\xi(\partial/\partial x^2)^a](\partial/\partial x^3)^b + gab(\partial/\partial x^2)^a\mathcal{L}_\xi(\partial/\partial x^3)^b = 0, \end{aligned}$$

where we used $\mathcal{L}_\xi(\partial/\partial x^2)^a = [\xi, \partial/\partial x^2]^a = [\partial/\partial t, \partial/\partial x^2]^a = 0$ and $\mathcal{L}_\xi(\partial/\partial x^3)^a = 0$ in the last step.

Now let $\Omega^2 \equiv g_{22}$, $\Lambda \equiv g_{33}/\Omega^2$, $w = W/V$, then (8.5.3) can be rewritten as

$$ds^2 = -V(dt - w d\varphi)^2 + V^{-1}\rho^2 d\varphi^2 + \Omega^2(d\rho^2 + \Lambda dz^2). \quad (8.5.5)$$

Thus, the number of functions of two variables that determine the components of the metric is reduced from 6 to 4, namely $V(\rho, z)$, $w(\rho, z)$, $\Omega(\rho, z)$ and $\Lambda(\rho, z)$. If the equation to be solved is the vacuum Einstein equation, then (8.5.5) can also be simplified as [see Wald (1984) p. 166]

⁵ This definition will be invalid when $\nabla_a\rho = 0$, and hence the coordinate patch does not contain points with $\nabla_a\rho = 0$.

$$ds^2 = -V(dt - w d\varphi)^2 + V^{-1}[\rho^2 d\varphi^2 + e^{2\gamma}(d\rho^2 + dz^2)], \quad \gamma \equiv \frac{1}{2} \ln(V\Omega^2). \quad (8.5.6)$$

The equation above indicates that the undetermined functions of two variables are now reduced from 4 to 3, namely $V(\rho, z)$, $w(\rho, z)$ and $\gamma(\rho, z)$. In the special case of $V = 1$, $w = \gamma = 0$, the equation above will turn into the line element expression of the Minkowski metric in the cylindrical coordinate system

$$ds^2 = -dt^2 + \rho^2 d\varphi^2 + d\rho^2 + dz^2.$$

Readers interested in the derivation of (8.5.6) may refer to Chap. 20 in Stephani et al. (2003), while those who only want to see the conclusion and a sketch of the derivation may refer to Wald (1984) pp. 166–168.

An important example of a stationary axisymmetric solution to the vacuum Einstein solution is the Kerr solution, which describes the exterior spacetime geometry of a particular kind of uncharged rotating star,⁶ see Chap. 13 for details.

If an axisymmetric metric also has translational invariance along the axis of symmetry, then it is called a **cylindrically symmetric metric**. Precisely speaking, besides the Killing vector field reflecting the axial symmetry, for a cylindrically symmetric metric there also exists a Killing vector field η^a reflecting the “translational invariance along the axis”, which satisfies ① $[\eta, \psi]^a = 0$; ② the integral curves of η^a are homeomorphic to \mathbb{R} .

Readers interested in cylindrically symmetric metrics may refer to Chap. 22 in Stephani et al. (2003).

8.6 Plane Symmetric Metrics [Optional Reading]

Before the definition of a spherically symmetric metric was given in Sect. 8.2, we have discussed the symmetry of a 2-dimensional surface (S^2, h_{ab}) in 3-dimensional Euclidean space. In a similar sense, we shall go over the symmetry of a 2-dimensional Euclidean plane $(\mathbb{R}^2, \delta_{ab})$ before introducing the definition of a plane symmetric metric. In a simple manner, we have found all 3 independent Killing vector fields of $(\mathbb{R}^2, \delta_{ab})$ in Example (1) of Sect. 4.3, i.e., $\xi_1^a \equiv (\partial/\partial x)^a$ and $\xi_2^a \equiv (\partial/\partial y)^a$ reflecting the translational invariance and $\xi_3^a \equiv -y(\partial/\partial x)^a + x(\partial/\partial y)^a$ reflecting the rotational invariance. From the linear combinations of $\xi_1^a, \xi_2^a, \xi_3^a$ one can have infinitely many Killing vector fields (note that the coefficients should be constants instead of functions on \mathbb{R}^2), and the corresponding isometries form a 3-parameter group of isometries, called the **Euclidean group**, denoted by $E(2)$ (see Sect. G.5.5 in Volume II for details). Following the definition of a spherically symmetric metric (see Definition 1 of Sect. 8.2), we have the following definition of a plane symmetric metric:

Definition 1 A spacetime metric g_{ab} is said to be **plane symmetric** if its group of isometries has a subgroup G_3 that is isomorphic to $E(2)$, and all the orbits of G_3 are 2-dimensional planes.

H. Taub proved the following theorem [Taub (1951)]: a plane symmetric solution to the vacuum Einstein equation must be a static metric, whose line element expression is

$$ds^2 = \frac{1}{\sqrt{1+kZ}}(-dT^2 + dZ^2) + (1+kZ)(dX^2 + dY^2), \quad (8.6.1)$$

⁶ Not the exterior spacetimes of all uncharged rotating stars can be described by the Kerr solution, see Hawking and Ellis (1973) p. 161 for this caveat.

where k is a constant. The coefficient of $(-dT^2 + dZ^2)$ being positive indicates that T and Z are respectively timelike and spacelike coordinates. The components of the metric not containing T means that $(\partial/\partial T)^a$ is a timelike Killing field, and thus the metric is static. At the beginning, Taub's paper only required the metric to have the plane symmetry, i.e., it only required three Killing vector fields $(\partial/\partial X)^a$, $(\partial/\partial Y)^a$ and $-Y(\partial/\partial X)^a + X(\partial/\partial Y)^a$, based on which he showed that it must contain the fourth (extra) Killing vector field $(\partial/\partial T)^a$. This is very much like Birkhoff's theorem. Moreover, Taub's original theorem has the same shortcoming as the Birkhoff's theorem: it omitted another possibility when deriving (8.6.1) which is on an equal footing with it. In fact, it can be proved from the vacuum condition and the plane symmetry that the metric will have either the form of (8.6.1) or the following form:

$$ds^2 = -\frac{1}{\sqrt{1+kZ}}(-dT^2 + dZ^2) + (1+kZ)(dX^2 + dY^2). \quad (8.6.2)$$

The coefficient of $(-dT^2 + dZ^2)$ in the above equation is negative, which means Z is a timelike coordinate and T is a spacelike coordinate. The metric components not depending on T indicates that $(\partial/\partial T)^a$ is a spacelike Killing field; together with the other two spatial Killing fields $(\partial/\partial X)^a$ and $(\partial/\partial Y)^a$, this indicates that the spacetime is **spatially homogeneous**, since it has the translational invariance in the three spatial directions (represented by the T -, X - and Y -axes). This metric does not have a timelike Killing vector field, and hence is not static. Thus, Taub's theorem should be revised as follows: a plane symmetric solution to the vacuum Einstein equation is either static or spatially homogeneous.

Another drawback of Taub's original paper is that (8.6.1) contains an arbitrary constant k , which may mislead people to think that (8.6.1), just like the Schwarzschild metric, is a one-parameter family of metrics. (Indeed, the parameter M of the Schwarzschild metric indicates that it is a one-parameter family). In the case $k \neq 0$, we introduce new coordinates $t = k^{-1/3}T$, $z = k^{-4/3}(1+kZ)$, $x = k^{2/3}X$ and $y = k^{2/3}Y$, then (8.6.1) and (8.6.2) will turn into

$$ds^2 = z^{-1/2}(-dt^2 + dz^2) + z(dx^2 + dy^2), \quad (8.6.1')$$

$$ds^2 = -z^{-1/2}(-dt^2 + dz^2) + z(dx^2 + dy^2). \quad (8.6.2')$$

This indicates that in the case $k \neq 0$, each of (8.6.1) and (8.6.2) represents one metric rather than a family of metrics. From this aspect, the Taub metric is very much different from the Schwarzschild metric.

The study of plane symmetric solutions of the electrovac Einstein equation can be dated back to 1926. However, the discovery of the general solution of this type started in the 1970s. Based on the work of Patnaik (1970), Letelier and Tabenski (1974) found the general solution of a plane symmetric metric produced by a plane symmetric electromagnetic field [see Stephani et al. (2003)]

$$ds^2 = \frac{1}{2}Y'(z)(-dt^2 + dz^2) + Y^2(z)(dx^2 + dy^2), \quad (8.6.3)$$

where $Y'(z) \equiv dY/dz$, and $Y(z)$ is given implicitly by the following equation:

$$(Y - A)^2 + 2A^2 \ln(Y + A) = -Cz, \quad A, C \text{ are constants.} \quad (8.6.4)$$

The electromagnetic field F_{ab} corresponding to (8.6.3) is a source-free nonnull electromagnetic field, whose coordinate components are (t, x, y, z are identified as x^0, x^1, x^2, x^3 , respectively)

$$F_{12} = C_1, \quad F_{30} = \frac{C_2}{2} Y' Y^{-2}, \quad A \equiv \frac{4\pi}{C} (C_1^2 + C_2^2), \quad C_1, C_2 \text{ are constants.} \quad (8.6.5)$$

When $F_{ab} = 0$, the metric (8.6.3) will be simplified to (8.6.1') [for $(\nabla_a Y) \nabla^a Y < 0$] or (8.6.2') [for $(\nabla_a Y) \nabla^a Y > 0$].

The expression (8.6.3) represents the plane symmetric metric produced by a plane symmetric electromagnetic field F_{ab} . The so-called plane symmetric electromagnetic field refers to

$$\mathcal{L}_{\xi_i} F_{ab} = 0, \quad i = 1, 2, 3, \quad (8.6.6)$$

where ξ_i^a represents the three Killing vector fields reflecting the plane symmetry, i.e.,

$$\xi_1^a \equiv (\partial/\partial x)^a, \quad \xi_2^a \equiv (\partial/\partial y)^a, \quad \xi_3^a \equiv -y(\partial/\partial x)^a + x(\partial/\partial y)^a. \quad (8.6.7)$$

It is not difficult to verify that (Exercise 8.9) the F_{ab} in (8.2.5) satisfies (8.6.6). However, a plane symmetric metric can also be produced by a non-plane symmetric electromagnetic field. An electromagnetic field with only translational symmetries but no rotational symmetry [i.e., (8.6.6) only holds for $i = 1, 2$] is called a **semi-plane symmetric** electromagnetic field ("2/3-plane symmetric" may be more appropriate). Some special solutions of a plane symmetric metric produced by this kind of electromagnetic field are scattered in the literature. Li and Liang (1985) found the general solutions of plane symmetric metrics produced by semi-plane symmetric electromagnetic fields, and classified them into two types:

$$\text{Type A} \quad ds^2 = \pm \frac{J(T+Z)}{\sqrt{T}} (-dT^2 + dZ^2) + T(dX^2 + dY^2), \quad (8.6.8a)$$

$$\text{Type B} \quad ds^2 = \pm \frac{J(T+Z)}{\sqrt{T+Z}} (-dT^2 + dZ^2) + (T+Z)(dX^2 + dY^2), \quad (8.6.8b)$$

where $J(T+Z)$ is an arbitrary function satisfying $\dot{J}/J > 0$ ($\dot{J} \equiv \partial J/\partial T$).⁷ The electromagnetic field corresponding to (8.6.8a) and (8.6.8b) is a semi- (2/3-) plane symmetric source-free null electromagnetic field. The general solutions (8.6.3) and (8.6.8) correspond to a nonnull, plane symmetric and a null, semi-plane symmetric source-free electromagnetic field, respectively. It is natural to ask: is there any plane symmetric metric produced by an electromagnetic field (no matter what symmetry it has) other than (8.6.3) and (8.6.8)? Kuang et al. (1987) proved that: ① The plane symmetric metrics produced by electromagnetic fields only have three types, namely (8.6.3), (8.6.8a) and (8.6.8b) (and the line elements obtained from them by coordinate transformations); ② The plane symmetric metric (8.6.3) cannot be produced by an electromagnetic field with source; ③ Plane symmetric metrics (8.6.8a) and (8.6.8b) can also be produced by electromagnetic fields with source, i.e., every metric of type A or B can be interpreted as either being produced by a source-free electromagnetic field or an electromagnetic field with source. [These two interpretations correspond to the same energy-momentum tensor T_{ab} , called a dual interpretation.⁸] Both of them are null electromagnetic fields; the former is semi- (2/3-) plane symmetric (has only translational

⁷ The line elements given by two different functions $J(T+Z)$ based on either (8.6.8a) or (8.6.8b) could differ only by a coordinate transformation (i.e., one can be obtained from the other via a coordinate transformation). Such two line elements represent the same geometry, and thus such two functions $J(T+Z)$ are said to be equivalent. To figure out all the different geometries described by (8.6.8a) and (8.6.8b), one needs to find the criterion for determining whether two arbitrary functions $J(T+Z)$ are equivalent. This necessary and sufficient criterion was found in Kuang et al. (1986).

⁸ The energy-momentum tensor T'_{ab} of the source of the electromagnetic field (dust) should also appear on the right-hand side of Einstein's equation just like the energy-momentum tensor T_{ab} of the electromagnetic field, which makes the question very complicated. One simplified discussion is to stipulate that $T'_{ab} = 0$, see Tariq and Tupper (1976) for its physical meaning.

symmetries but no rotational symmetry), while the latter one, on the contrary, has only rotational symmetry but no translational symmetry (i.e., $\mathcal{L}_{\xi_3} F_{ab} = 0$, $\mathcal{L}_{\xi_1} F_{ab} \neq 0$, $\mathcal{L}_{\xi_2} F_{ab} \neq 0$), which may also be called a semi-plane symmetric electromagnetic field (of another kind), or more precisely a 1/3-plane symmetric electromagnetic field. With this, the plane symmetric metrics produced by electromagnetic fields are finally exhausted.

The fact that a plane symmetric metric can be produced by a semi-plane symmetric electromagnetic field indicates that the symmetry of the electromagnetic field can be weaker than the symmetry of the metric. It is natural to ask: can the symmetry of the metric be weaker than the symmetry of the electromagnetic field? For example, does there exist a semi-plane symmetric metric produced by a plane symmetric electromagnetic field? The answer is affirmative: Li and Liang (1989) provided a specific example (a special solution).

We mention in passing that the three Killing fields reflecting the spherical symmetry are on an equal footing, and there does not exist any spherically symmetric metric produced by a semi-(2/3- or 1/3-) spherically symmetric electromagnetic field. A spherically symmetric metric produced by an electromagnetic field can only be the RN metric, whose electromagnetic field can only be a spherically symmetric, source-free nonnull electromagnetic field.

8.7 The Newman-Penrose (NP) Formalism [Optional Reading]

Besides the coordinate basis method and the orthonormal tetrad method, there is also a third commonly used method of computing curvature, that is the “null tetrad method” proposed by Newman and Penrose (1962). This method can be viewed as a variant of the rigid tetrad method: instead of using an orthonormal tetrad, here one uses a complex⁹ “null tetrad”. Suppose p is a point of a 4-dimensional spacetime (M, g_{ab}) , and $\{(e_\mu)^a\}$ is an orthonormal tetrad at p . Define 4 special vectors at p as follows:

$$\begin{aligned} m^a &:= \frac{1}{\sqrt{2}}[(e_1)^a - i(e_2)^a], & \bar{m}^a &:= \frac{1}{\sqrt{2}}[(e_1)^a + i(e_2)^a], \\ l^a &:= \frac{1}{\sqrt{2}}[(e_0)^a - (e_3)^a], & k^a &:= \frac{1}{\sqrt{2}}[(e_0)^a + (e_3)^a], \end{aligned} \quad (8.7.1)$$

then $g_{ab}m^a m^b = g_{ab}\bar{m}^a \bar{m}^b = g_{ab}l^a l^b = g_{ab}k^a k^b = 0$, i.e., all 4 of them are null vectors. Note that m^a and \bar{m}^a are both complex vectors conjugate to each other. To distinguish from other tetrads, this text will use $\{(\varepsilon_\mu)^a\}$ to represent a null tetrad, and stipulate the numbering as [in agreement with Stephani et al. (2003)]

$$(\varepsilon_1)^a \equiv m^a, \quad (\varepsilon_2)^a \equiv \bar{m}^a, \quad (\varepsilon_3)^a \equiv l^a, \quad (\varepsilon_4)^a \equiv k^a. \quad (8.7.2)$$

The corresponding dual basis vectors are

$$(\varepsilon^1)_a \equiv \bar{m}_a, \quad (\varepsilon^2)_a \equiv m_a, \quad (\varepsilon^3)_a \equiv -k_a, \quad (\varepsilon^4)_a \equiv -l_a. \quad (8.7.2')$$

⁹ Change the \mathbb{R} in Definition 2 of Sect. 2.2 to \mathbb{C} , then a map $v : \mathcal{F}_M \rightarrow \mathbb{C}$ is called a complex vector, and thus the tangent space V_p at p is generalized to an n -dimensional complex vector space (the scalar multiplication uses complex numbers). Suppose u and w are real vectors at p and $v(f) = u(f) + iw(f)$, $\forall f \in \mathcal{F}_M$, then we say that $v = u + iw$, and call u and w the real and imaginary parts of v , respectively. Similarly, it is not difficult to define a complex tensor as well as its real and imaginary parts.

$(\varepsilon_\mu)^a$ can be regarded as a special case of an arbitrary basis field $(e_\mu)^a$, which we mentioned at the beginning of Sect. 5.7; however, one should not confuse this with the $(e_\mu)^a$ in (8.7.1), which only refers to an orthonormal tetrad. It is not difficult to see that the inner product of any two basis vectors in a null tetrad has only the following two pairs of nonzero ones:

$$m^a \bar{m}_a \equiv g_{ab} m^a \bar{m}^b = g_{12} = g_{21} = 1, \quad l^a k_a \equiv g_{ab} l^a k^b = g_{34} = g_{43} = -1,$$

and thus the matrices constituted by the components $g_{\mu\nu}$ and $g^{\mu\nu}$ of the metric g_{ab} and its inverse g^{ab} are

$$(g_{\mu\nu}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix} = (g^{\mu\nu}). \quad (8.7.3)$$

Just like in §5.7, the number indices μ of $(\varepsilon_\mu)^a$ and $(\varepsilon^\mu)_a$ can also be raised and lowered using $g^{\mu\nu}$ and $g_{\mu\nu}$. Applying (5.7.5) to a null tetrad yields

$$\omega_\mu{}^v{}_a = (\varepsilon_\mu)^c \nabla_a (\varepsilon^v)_c, \quad (8.7.4)$$

and the corresponding Ricci rotation coefficients are

$$\omega_\mu{}^v{}_\rho = (\varepsilon_\mu)^b (\varepsilon_\rho)^a \nabla_a (\varepsilon^v)_b.$$

Equation (8.7.3) indicates that $(\varepsilon_\mu)^a$ is a (complex) rigid tetrad, and hence we have $\omega_{\mu\nu a} = (\varepsilon_\mu)_b \nabla_a (\varepsilon_\nu)^b$ and $\omega_{\mu\nu a} = -\omega_{\nu\mu a}$ (i.e., $\omega_{\mu\nu} = -\omega_{\nu\mu}$), and for the corresponding Ricci rotation coefficients

$$\omega_{\mu\nu\rho} = (\varepsilon_\mu)^b (\varepsilon_\rho)^a \nabla_a (\varepsilon_\nu)_b, \quad \omega_{\mu\nu\rho} = -\omega_{\nu\mu\rho}. \quad (8.7.5)$$

Since the numbering of the null tetrad indices are 1, 2, 3, 4 rather than 0, 1, 2, 3, the number indices of the corresponding connection 1-forms are also changed to $\omega_{12}, \omega_{13}, \omega_{14}, \omega_{23}, \omega_{24}, \omega_{34}$. Note that $\omega_{\mu\nu\rho}$ corresponding to a null tetrad has a complex value, which obeys the following proposition:

Proposition 8.7.1 *If we exchange all the 1s and 2s in the subscripts of $\omega_{\mu\nu\rho}$ (and keep all the 3s and 4s unchanged), we obtain its complex conjugate $\bar{\omega}_{\mu\nu\rho}$, e.g., $\omega_{134} = \bar{\omega}_{234}$, $\omega_{342} = \bar{\omega}_{341}$, $\omega_{421} = \bar{\omega}_{412}$, $\omega_{122} = \bar{\omega}_{211}$, $\omega_{344} = \bar{\omega}_{344}$.*

Proof It follows from (8.7.5) that $\bar{\omega}_{\mu\nu\rho} = (\bar{\varepsilon}_\mu)^b (\bar{\varepsilon}_\rho)^a \nabla_a (\bar{\varepsilon}_\nu)_b$, and it is not difficult to prove this proposition using this equation. For example,

$$\bar{\omega}_{412} = (\bar{\varepsilon}_4)^b (\bar{\varepsilon}_2)^a \nabla_a (\bar{\varepsilon}_1)_b = (\varepsilon_4)^b (\varepsilon_1)^a \nabla_a (\varepsilon_2)_b = \omega_{421}. \quad \square$$

Proposition 8.7.1 not only holds for $\omega_{\mu\nu\rho}$, but also holds for all the quantities (including tensors) that carry null tetrad indices, e.g., $\omega_{41} = \bar{\omega}_{42}$, $\omega_{21} = \bar{\omega}_{12}$, $\mathbf{R}_{31} = \bar{\mathbf{R}}_{32}$, $\mathbf{R}_{12} = \bar{\mathbf{R}}_{21}$, $\mathbf{R}_{34} = \bar{\mathbf{R}}_{34}$.

The process of computing the curvature tensor using the null tetrad method is similar to that using the orthonormal tetrad method; that is, one finds all the connection 1-forms $\omega_{\mu\nu}$ of the chosen null tetrad and then finds all the curvature 2-forms $\mathbf{R}_{\mu\nu}$. The components $\omega_{\mu\nu\rho}$ of the connection 1-forms can still be computed from (5.7.19) and (5.7.20), in which the $(e_\mu)^a$ should now be interpreted as $(\varepsilon_\mu)^a$. After finding all the $\omega_{\mu\nu}$ one can still use Cartan's second equation of structure to compute all the $\mathbf{R}_{\mu\nu}$.

Proposition 8.7.2 *In a null tetrad, Cartan's second structure equation (5.7.8) reads*

$$\mathbf{R}_{41} = d\omega_{41} + \omega_{41} \wedge (\omega_{21} + \omega_{43}), \quad (8.7.6a)$$

$$\mathbf{R}_{32} = d\omega_{32} - \omega_{32} \wedge (\omega_{21} + \omega_{43}), \quad (8.7.6b)$$

$$\mathbf{R}_{21} + \mathbf{R}_{43} = d(\omega_{21} + \omega_{43}) + 2\omega_{32} \wedge \omega_{41}. \quad (8.7.6c)$$

Proof When we have a metric g_{ab} , Cartan's second equation (5.7.8) can be written as

$$R_{\mu\nu} = d\omega_{\mu\nu} + \omega_\mu{}^\tau \wedge \omega_{\tau\nu} = d\omega_{\mu\nu} + g^{\lambda\tau} \omega_{\mu\lambda} \wedge \omega_{\tau\nu},$$

where $g^{\lambda\tau}$ are the components of g^{ab} in the null tetrad. Noticing that the only nonzero $g^{\lambda\tau}$ are $g^{12} = g^{21} = 1$ and $g^{34} = g^{43} = -1$, we can write down all 6 independent components of $\mathbf{R}_{\mu\nu}$ as follows:

$$\mathbf{R}_{43} = d\omega_{43} + \omega_{41} \wedge \omega_{23} + \omega_{42} \wedge \omega_{13}, \quad (8.7.7a)$$

$$\mathbf{R}_{42} = d\omega_{42} + \omega_{42} \wedge (\omega_{12} + \omega_{43}), \quad (8.7.7b)$$

$$\mathbf{R}_{41} = d\omega_{41} + \omega_{41} \wedge (\omega_{21} + \omega_{43}), \quad (8.7.7c)$$

$$\mathbf{R}_{32} = d\omega_{32} + \omega_{32} \wedge (\omega_{12} + \omega_{34}), \quad (8.7.7d)$$

$$\mathbf{R}_{31} = d\omega_{31} + \omega_{31} \wedge (\omega_{21} + \omega_{34}), \quad (8.7.7e)$$

$$\mathbf{R}_{21} = d\omega_{21} - \omega_{23} \wedge \omega_{41} - \omega_{24} \wedge \omega_{31}. \quad (8.7.7f)$$

Considering Proposition 8.7.1, these 6 equalities are not all independent, since from $\mathbf{R}_{31} = \bar{\mathbf{R}}_{32}$ and $\mathbf{R}_{42} = \bar{\mathbf{R}}_{41}$ we can derive (8.7.7e) and (8.7.7b) from (8.7.7d) and (8.7.7c). Moreover, (8.7.7a) and (8.7.7f) can be written as

$$\mathbf{R}_{43} = d\omega_{43} + \omega_{32} \wedge \omega_{41} + \overline{\omega_{32} \wedge \omega_{41}} = d\omega_{43} + 2\operatorname{Re}(\omega_{32} \wedge \omega_{41}), \quad (8.7.7a')$$

$$\mathbf{R}_{21} = d\omega_{21} + \omega_{32} \wedge \omega_{41} - \overline{\omega_{32} \wedge \omega_{41}} = d\omega_{21} + 2i\operatorname{Im}(\omega_{32} \wedge \omega_{41}), \quad (8.7.7f')$$

These two equations together are equivalent to (8.7.6c). Therefore, (8.7.7a)–(8.7.7f) are equivalent to (8.7.6a)–(8.7.6c). \square

The whole formalism introduced by Newman and Penrose based on the null tetrad method is called the **Newman-Penrose formalism**, or **NP formalism** for short. The basic idea of the NP formalism is to separate all kinds of sets of condensed equations [e.g., (8.7.6)] into multiple component equations, which will certainly lead to the appearance of quantities with many indices, such as $\omega_{\mu\nu\rho}$, $R_{\rho\sigma\mu\nu}$, etc. For the sake of making the equations look simpler (and other purposes), the NP formalism uses many notations which carry less or no indices to represent these quantities with many indices. We will introduce all three kinds of them as follows:

(1) Due to Proposition 8.7.1, only 12 out of the 24 linear combinations of the complex $\omega_{\mu\nu\rho}$ are linearly independent. (Comparing with the fact that there are 24 linearly independent real $\omega_{\mu\nu\rho}$ in a orthonormal tetrad, you will find this is quite natural). Use 12 Greek letters without indices to represent 12 linearly independent combinations of $\omega_{\mu\nu\rho}$ as follows [(8.7.5) is used]:

$$\kappa \equiv -\omega_{144} = -m^a k^b \nabla_b k_a, \quad (8.7.8a)$$

$$\rho \equiv -\omega_{142} = -m^a \bar{m}^b \nabla_b k_a, \quad (8.7.8b)$$

$$\sigma \equiv -\omega_{141} = -m^a m^b \nabla_b k_a, \quad (8.7.8c)$$

$$\tau \equiv -\omega_{143} = -m^a l^b \nabla_b k_a, \quad (8.7.8d)$$

$$\nu \equiv \omega_{233} = \bar{m}^a l^b \nabla_b l_a, \quad (8.7.8e)$$

$$\mu \equiv \omega_{231} = \bar{m}^a m^b \nabla_b l_a , \quad (8.7.8f)$$

$$\lambda \equiv \omega_{232} = \bar{m}^a \bar{m}^b \nabla_b l_a , \quad (8.7.8g)$$

$$\pi \equiv \omega_{234} = \bar{m}^a k^b \nabla_b l_a , \quad (8.7.8h)$$

$$\varepsilon \equiv \frac{1}{2}(\omega_{214} - \omega_{344}) = \frac{1}{2}(\bar{m}^a k^b \nabla_b m_a - l^a k^b \nabla_b k_a) , \quad (8.7.8i)$$

$$\beta \equiv \frac{1}{2}(\omega_{211} - \omega_{341}) = \frac{1}{2}(\bar{m}^a m^b \nabla_b m_a - l^a m^b \nabla_b k_a) , \quad (8.7.8j)$$

$$\gamma \equiv \frac{1}{2}(\omega_{433} - \omega_{123}) = \frac{1}{2}(k^a l^b \nabla_b l_a - m^a l^b \nabla_b \bar{m}_a) , \quad (8.7.8k)$$

$$\alpha \equiv \frac{1}{2}(\omega_{432} - \omega_{122}) = \frac{1}{2}(k^a \bar{m}^b \nabla_b l_a - m^a \bar{m}^b \nabla_b \bar{m}_a) . \quad (8.7.8l)$$

These 12 greek letters are called the **spin coefficients**.

Proposition 8.7.3 *The 24 $\omega_{\mu\nu\rho}$ can be expressed in terms of the 12 spin coefficients as follows:*

$$\begin{aligned} \omega_{121} &= \bar{\alpha} - \beta , & \omega_{122} &= \bar{\beta} - \alpha , & \omega_{123} &= \bar{\gamma} - \gamma , & \omega_{124} &= \bar{\varepsilon} - \varepsilon , \\ \omega_{131} &= \bar{\lambda} , & \omega_{132} &= \bar{\mu} , & \omega_{133} &= \bar{\nu} , & \omega_{134} &= \bar{\pi} , \\ \omega_{141} &= -\sigma , & \omega_{142} &= -\rho , & \omega_{143} &= -\tau , & \omega_{144} &= -\kappa , \\ \omega_{231} &= \mu , & \omega_{232} &= \lambda , & \omega_{233} &= \nu , & \omega_{234} &= \pi , \\ \omega_{241} &= -\bar{\rho} , & \omega_{242} &= -\bar{\sigma} , & \omega_{243} &= -\bar{\tau} , & \omega_{244} &= -\bar{\kappa} , \\ \omega_{341} &= -(\bar{\alpha} + \beta) , & \omega_{342} &= -(\alpha + \bar{\beta}) , & \omega_{343} &= -(\gamma + \bar{\gamma}) , & \omega_{344} &= -(\varepsilon + \bar{\varepsilon}) . \end{aligned}$$

Proof Only 8 out of the 24 equations above need to be checked (the others can be read directly from the definition of the spin coefficients), the verification is as follows.

Firstly, since $(\varepsilon_3)^a$ and $(\varepsilon_4)^a$ are real vectors, ω_{343} and ω_{344} are real. Secondly, it follows from $\omega_{213} = -\omega_{123} = -\bar{\omega}_{213}$ that $\omega_{213} + \bar{\omega}_{213} = 0$, and hence ω_{213} is imaginary. Similarly we can see that ω_{214} is also imaginary. Also, $\varepsilon \equiv \frac{1}{2}(\omega_{214} - \omega_{344}) = \frac{1}{2}(\omega_{434} + \omega_{214})$, and hence $\omega_{434} = 2\text{Re}(\varepsilon) = \varepsilon + \bar{\varepsilon}$, $\omega_{214} = 2\text{Im}(\varepsilon) = \varepsilon - \bar{\varepsilon}$. Similarly, we have $\omega_{433} = \gamma + \bar{\gamma}$, $\omega_{213} = \gamma - \bar{\gamma}$. Furthermore, from the definitions of α and β we get $\beta = -\frac{1}{2}(\omega_{121} + \omega_{341})$, $\bar{\alpha} = \frac{1}{2}(\omega_{121} - \omega_{341})$. Thus, $\omega_{341} = -(\bar{\alpha} + \beta)$, $\omega_{121} = \bar{\alpha} - \beta$, from which we can easily get $\omega_{122} = \bar{\beta} - \alpha$, $\omega_{342} = -(\alpha + \bar{\beta})$. \square

(2) Since the derivatives of spin coefficients along the 4 basis vectors appear frequently in all kinds of equations, we introduce the following 4 notations for derivatives:

$$\delta \equiv m^a \nabla_a , \quad \bar{\delta} \equiv \bar{m}^a \nabla_a , \quad \Delta \equiv l^a \nabla_a , \quad D \equiv k^a \nabla_a . \quad (8.7.9)$$

(3) The components of the Riemann tensor $R_{abc}{}^d$ have 4 indices. We would like to denote them using notations with less indices. $R_{abc}{}^d$ is determined by its “traceless part” (Weyl tensor) $C_{abc}{}^d$ and “trace part” (Ricci tensor) R_{ab} . Due to various symmetries, the Weyl tensor has only 10 real independent components, which can be represented by 5 complex quantities $\Psi_0, \Psi_1, \Psi_2, \Psi_3, \Psi_4$ defined as

$$\begin{aligned} \Psi_0 &:= C_{4141} , & \Psi_1 &:= C_{4341} , & \Psi_2 &:= \frac{1}{2}(C_{4343} - C_{4312}) , \\ & & & & \Psi_3 &:= C_{3432} , & \Psi_4 &:= C_{3232} , \end{aligned} \quad (8.7.10)$$

where $C_{\mu\nu\rho\sigma}$ are the components of C_{abcd} in the null tetrad. The Ricci tensor R_{ab} only has 10 real independent components due to the symmetry $R_{ab} = R_{ba}$. In the null tetrad, among

the 10 independent components $R_{44}, R_{43}, R_{42}, R_{41}, R_{33}, R_{32}, R_{31}, R_{22}, R_{21}, R_{11}$, 6 are complex and 4 are real. It is obvious that R_{44}, R_{43}, R_{33} are real, and R_{21} is also real since $R_{21} = R_{12} = \bar{R}_{21}$. In terms of linear combinations of these 4 real numbers, one can define the following 4 real quantities:

$$\Phi_{00} := \frac{1}{2}R_{44}, \quad \Phi_{11} := \frac{1}{4}(R_{21} + R_{43}), \quad \Phi_{22} := \frac{1}{2}R_{33}, \quad R := 2(R_{21} - R_{43}). \quad (8.7.11a)$$

The fourth real quantity R is actually the scalar curvature [it is easy to show that the scalar curvature indeed equals $2(R_{21} - R_{43})$]. In terms of the 6 complex components $R_{42}, R_{41}, R_{32}, R_{31}, R_{22}, R_{11}$, one can define 6 complex quantities

$$\begin{aligned} \Phi_{01} &:= \frac{1}{2}R_{41}, & \Phi_{10} &:= \frac{1}{2}R_{42}, & \Phi_{02} &:= \frac{1}{2}R_{11}, \\ \Phi_{20} &:= \frac{1}{2}R_{22}, & \Phi_{12} &:= \frac{1}{2}R_{31}, & \Phi_{21} &:= \frac{1}{2}R_{32}. \end{aligned} \quad (8.7.11b)$$

The above 10 quantities excluding R can be arranged into a 3×3 “conjugate symmetric” matrix $[\Phi_{\lambda\tau}]$ (satisfying $\Phi_{\lambda\tau} = \bar{\Phi}_{\tau\lambda}$, $\lambda, \tau = 0, 1, 2$):

	0	1	2
0	$\frac{1}{2}R_{44}$	$\frac{1}{2}R_{41}$	$\frac{1}{2}R_{11}$
1	$\frac{1}{2}R_{42}$	$\frac{1}{4}(R_{21} + R_{43})$	$\frac{1}{2}R_{31}$
2	$\frac{1}{2}R_{22}$	$\frac{1}{2}R_{32}$	$\frac{1}{2}R_{33}$

The 3 independent off-diagonal elements together with the 3 real diagonal elements and the real number R represent exactly the 10 real independent components of R_{ab} .

The NP formalism contains 3 equation systems that are very useful, namely (A) the NP equations; (B) the Bianchi identities; (C) the commutation relations. Here we introduce them as follows.

(A) NP equations.

Expressing $\mathbf{R}_{41}, \mathbf{R}_{32}, \mathbf{R}_{21}, \mathbf{R}_{43}$ in terms of $\Psi_0, \Psi_1, \Psi_2, \Psi_3, \Psi_4$ as well as the 10 quantities $\Phi_{00}, \dots, \Phi_{22}$ and R , and expressing $\omega_{41}, \omega_{32}, \omega_{21}, \omega_{43}$ in terms of the 12 spin coefficients, one can reformulate (8.7.6) into the following 18 equations, called the **NP equations**:

$$D\rho - \bar{\delta}\kappa = (\rho^2 + \sigma\bar{\sigma}) + \rho(\varepsilon + \bar{\varepsilon}) - \bar{\kappa}\tau - \kappa(3\alpha + \bar{\beta} - \pi) + \Phi_{00}, \quad (8.7.12a)$$

$$D\sigma - \delta\kappa = \sigma(\rho + \bar{\rho}) + \sigma(3\varepsilon - \bar{\varepsilon}) - \kappa(\tau - \bar{\pi} + \bar{\alpha} + 3\beta) + \Psi_0, \quad (8.7.12b)$$

$$D\tau - \Delta\kappa = \rho(\tau + \bar{\pi}) + \sigma(\bar{\tau} + \pi) + \tau(\varepsilon - \bar{\varepsilon}) - \kappa(3\gamma + \bar{\gamma}) + \Psi_1 + \Phi_{01}, \quad (8.7.12c)$$

$$D\alpha - \bar{\delta}\varepsilon = \alpha(\rho + \bar{\varepsilon} - 2\varepsilon) + \beta\bar{\sigma} - \bar{\beta}\varepsilon - \kappa\lambda - \bar{\kappa}\gamma + \pi(\varepsilon + \rho) + \Phi_{10}, \quad (8.7.12d)$$

$$D\beta - \delta\varepsilon = \sigma(\alpha + \pi) + \beta(\bar{\rho} - \bar{\varepsilon}) - \kappa(\mu + \gamma) - \varepsilon(\bar{\alpha} - \bar{\pi}) + \Psi_1, \quad (8.7.12e)$$

$$D\gamma - \Delta\varepsilon = \alpha(\tau + \bar{\pi}) + \beta(\bar{\tau} + \pi) - \gamma(\varepsilon + \bar{\varepsilon}) - \varepsilon(\gamma + \bar{\gamma}) + \tau\pi - \nu\kappa + \Psi_2 + \Phi_{11} - R/24, \quad (8.7.12f)$$

$$D\lambda - \bar{\delta}\pi = (\rho\lambda + \bar{\sigma}\mu) + \pi^2 + \pi(\alpha - \bar{\beta}) - \nu\bar{\kappa} - \lambda(3\varepsilon - \bar{\varepsilon}) + \Phi_{20}, \quad (8.7.12g)$$

$$D\mu - \delta\pi = (\bar{\rho}\mu + \sigma\lambda) + \pi\bar{\pi} - \mu(\varepsilon + \bar{\varepsilon}) - \pi(\bar{\alpha} - \beta) - \nu\kappa + \Psi_2 + R/12, \quad (8.7.12h)$$

$$D\nu - \Delta\pi = \mu(\pi + \bar{\tau}) + \lambda(\bar{\pi} + \tau) + \pi(\gamma - \bar{\gamma}) - \nu(3\varepsilon + \bar{\varepsilon}) + \psi_3 + \Phi_{21}, \quad (8.7.12i)$$

$$\Delta\lambda - \bar{\delta}\nu = -\lambda(\mu + \bar{\mu}) - \lambda(3\gamma - \bar{\gamma}) + \nu(3\alpha + \bar{\beta} + \pi - \bar{\tau}) - \Psi_4, \quad (8.7.12j)$$

$$\delta\rho - \bar{\delta}\sigma = \rho(\bar{\alpha} + \beta) - \sigma(3\alpha - \bar{\beta}) + \tau(\rho - \bar{\rho}) + \kappa(\mu - \bar{\mu}) - \Psi_1 + \Phi_{01}, \quad (8.7.12k)$$

$$\begin{aligned} \delta\alpha - \bar{\delta}\beta &= (\mu\rho - \lambda\sigma) + \alpha\bar{\alpha} + \beta\bar{\beta} - 2\alpha\beta + \gamma(\rho - \bar{\rho}) + \varepsilon(\mu - \bar{\mu}) \\ &\quad - \Psi_2 + \Phi_{11} + R/24, \end{aligned} \quad (8.7.12l)$$

$$\delta\lambda - \bar{\delta}\mu = \nu(\rho - \bar{\rho}) + \pi(\mu - \bar{\mu}) + \mu(\alpha + \bar{\beta}) + \lambda(\bar{\alpha} - 3\beta) - \Psi_3 + \Phi_{21}, \quad (8.7.12m)$$

$$\Delta\nu - \Delta\mu = (\mu^2 + \lambda\bar{\lambda}) + \mu(\gamma + \bar{\gamma}) - \bar{\nu}\pi + \nu(\tau - 3\beta - \bar{\alpha}) + \Phi_{22}, \quad (8.7.12n)$$

$$\delta\gamma - \Delta\beta = \gamma(\tau - \bar{\alpha} - \beta) + \mu\tau - \sigma\nu - \varepsilon\bar{\nu} - \beta(\gamma - \bar{\gamma} - \mu) + \alpha\bar{\lambda} + \Phi_{12}, \quad (8.7.12o)$$

$$\delta\tau - \Delta\sigma = (\mu\sigma + \bar{\lambda}\rho) + \tau(\tau + \beta - \bar{\alpha}) - \sigma(3\gamma - \bar{\gamma}) - \kappa\bar{\nu} + \Phi_{02}, \quad (8.7.12p)$$

$$\Delta\rho - \bar{\delta}\tau = -(\rho\bar{\mu} + \sigma\lambda) + \tau(\bar{\beta} - \alpha - \bar{\tau}) + \rho(\gamma + \bar{\gamma}) + \nu\kappa - \Psi_2 - R/12, \quad (8.7.12q)$$

$$\Delta\alpha - \bar{\delta}\gamma = \nu(\rho + \varepsilon) - \lambda(\tau + \beta) + \alpha(\bar{\gamma} - \bar{\mu}) + \gamma(\bar{\beta} - \bar{\tau}) - \Psi_3. \quad (8.7.12r)$$

Remark 1 Cartan's second equations (8.7.6) contain 3 equations of complex antisymmetric tensors of type $(0, 2)$, each of which is equivalent to 6 complex component equations, and thus there are 18 complex NP equations altogether.

Now we will illustrate the verification of the NP equations by some examples. First take (8.7.12a) as an example, it is in fact a reformulation of the fourth and second components of (8.7.6a). In the null tetrad, the components R_{4241} of R_{abcd} can be expressed as

$$R_{4241} = (\varepsilon_4)^a(\varepsilon_2)^b R_{ab41} = (\varepsilon_4)^a(\varepsilon_2)^b [(\mathrm{d}\omega_{41})_{ab} + \omega_{41a} \wedge (\omega_{21b} + \omega_{43b})],$$

where (8.7.6a) is used in the second step. Since $(\omega_{41})_b = \sigma(\varepsilon^1)_b + \rho(\varepsilon^2)_b + \tau(\varepsilon^3)_b + \kappa(\varepsilon^4)_b$, we have

$$\begin{aligned} (\varepsilon_4)^a(\varepsilon_2)^b (\mathrm{d}\omega_{41})_{ab} &= (\varepsilon_4)^a(\varepsilon_2)^b (\nabla_a \omega_{41b} - \nabla_b \omega_{41a}) \\ &= -\sigma\bar{\sigma} + \rho(\varepsilon - \bar{\varepsilon}) + D\rho - \rho^2 + \bar{\kappa}\tau - \kappa\pi + \kappa(\alpha + \bar{\beta}) - \bar{\delta}\kappa. \end{aligned}$$

The last step is tedious but not difficult, which is left as an exercise. The operation of lowering the index of $\omega_\mu{}^\nu{}_\rho$ occurs a lot in the derivation, which relies on the expression (8.7.3) for the components $g^{v\sigma}$ of g^{ab} in the null tetrad. Since the matrix in (8.7.3) is quite simple, it is pretty easy to do the calculation. For instance,

$$\omega_4{}^1{}_2 = g^{1\mu}\omega_{4\mu 2} = g^{12}\omega_{422} = \omega_{422}.$$

Moreover,

$$(\varepsilon_4)^a(\varepsilon_2)^b [\omega_{41a} \wedge (\omega_{21b} + \omega_{43b})] = \kappa(\omega_{212} + \omega_{432}) - \rho(\omega_{214} + \omega_{434}) = 2\kappa\alpha - 2\rho\varepsilon,$$

and hence

$$R_{4241} = (D\rho - \bar{\delta}\kappa) - (\rho^2 + \sigma\bar{\sigma}) - \rho(\varepsilon + \bar{\varepsilon}) + \bar{\kappa}\tau + \kappa(2\alpha + \bar{\beta} - \pi). \quad (8.7.13)$$

On the other hand, it follows from the definition of Φ_{00} and $R_{\mu\nu} = R_{\mu\sigma\nu}{}^\sigma$ that

$$\begin{aligned} \Phi_{00} &\equiv \frac{1}{2}R_{44} = \frac{1}{2}R_{4\mu 4}{}^\mu = \frac{1}{2}(R_{414}{}^1 + R_{424}{}^2 + R_{434}{}^3) \\ &= \frac{1}{2}(R_{4142} + R_{4241} - R_{4344}) = R_{4241}. \end{aligned} \quad (8.7.14)$$

Comparing (8.7.13) and (8.7.14) yields (8.7.12a). Thus, (8.7.12a) is nothing but a component equation of (8.7.6a). This might be unapparent for the beginning readers to see since Φ_{00} , which represents the curvature component, is written on the right-hand side of the equation. Now we introduce the derivation of a more complicated equation (8.7.12f). This is a reformulation of the 4th and 3rd component equations of (8.7.6c). First,

$$R_{4321} + R_{4343} = (\varepsilon_4)^a (\varepsilon_3)^b (R_{ab21} + R_{ab43}) = (\varepsilon_4)^a (\varepsilon_3)^b [(\mathrm{d}\omega_{21})_{ab} + (\mathrm{d}\omega_{43})_{ab} + 2\omega_{32a} \wedge \omega_{41b}],$$

where (8.7.6c) is used in the second equality. Through a tedious but straightforward computation we get

$$R_{4321} + R_{4343} = 2[(\mathrm{D}\gamma - \Delta\varepsilon) - \alpha(\tau + \bar{\pi}) - \beta(\bar{\tau} + \pi) + \gamma(\varepsilon + \bar{\varepsilon}) + \varepsilon(\gamma + \bar{\gamma}) - \tau\pi + \nu\kappa]. \quad (8.7.15)$$

On the other hand, from the definition (8.7.10) we know that $\Psi_2 = (C_{4343} - C_{4312})/2$. Applying the definition of the Weyl tensor [Equation (3.4.14)] to the $n = 4$ case yields

$$C_{abcd} = R_{abcd} - \frac{1}{2}[(g_{ac}R_{db} - g_{ad}R_{cb}) - (g_{bc}R_{da} - g_{bd}R_{ca})] + \frac{1}{6}R(g_{ac}g_{db} - g_{ad}g_{cb}).$$

Noticing (8.7.3), we have $C_{4343} = R_{4343} - R_{34} - R/6$, $C_{4312} = R_{4312}$, and hence

$$\Psi_2 = \frac{1}{2}(R_{4343} - R_{4312}) - \frac{1}{2}R_{34} - \frac{1}{12}R. \quad (8.7.16)$$

It follows from (8.7.10) that $\Phi_{11} = (R_{12} + R_{43})/4$ and $R = 2(R_{12} - R_{34})$, and hence

$$2(\Psi_2 + \Phi_{11} - R/24) = R_{4343} - R_{4312} = R_{4321} + R_{4343}. \quad (8.7.17)$$

From (8.7.17) and (8.7.15) we arrive at (8.7.12f).

In a word, the 18 NP equations are nothing but the manifestation of Cartan's second equation of structure in the null tetrad. One of their features is that the summations in the condensed expression (8.7.6) are listed one by one, which is convenient for practical computation. Although the set of NP equations contains a lot of equations, each of them involves only first order derivatives, and thus they are not so difficult to solve. By dint of the gauge freedom of choosing the null tetrad [there are 6 real parameters to choose, see Stephani et al. (2003) p. 33] one can even further simplify the NP equations.

(B) Bianchi identities.

From the definition of the Riemann tensor $R_{abc}{}^d$, we have already proved in Chap. 3 that it satisfies the Bianchi identity $\nabla_{[a}R_{bc]d}{}^e = 0$. For the convenience of application, one can formulate it into components equations by means of the NP null tetrad, see Stephani et al. (2003) pp. 81–82.

(C) Commutation relations.

To compute the Riemann tensor we need to choose a basis field $\{(e_\mu)^a\}$ first. If we choose the coordinate basis, then any two basis vector fields must commute with each other, i.e., $[\partial/\partial x^\mu, \partial/\partial x^\nu]^a = 0$. However, it is not as simple for a non-coordinate basis. The commutator of two arbitrary basis vector fields $(e_\mu)^a$ and $(e_\nu)^a$ in the basis $\{(e_\mu)^a\}$ can be expressed using (3.1.13) as

$$[e_\mu, e_\nu]^a = (e_\mu)^b \nabla_b (e_\nu)^a - (e_\nu)^b \nabla_b (e_\mu)^a, \quad (8.7.18)$$

where ∇_a is an arbitrary torsion-free derivative operator. Choose the derivative operator (connection) we assigned when computing the Riemann tensor as the ∇_a in the above equation, then it follows from (5.7.1) that

$$[e_\mu, e_\nu]^a = -2\gamma^\sigma_{[\mu\nu]}(e_\sigma)^a, \quad (8.7.19)$$

where $\gamma^\sigma_{\mu\nu}$ are the connection coefficients defined by (5.7.1), whose relation with the connection 1-form $\omega_\mu{}^\nu_a$ is given by (5.7.4). Equation (8.7.19) is exactly the **commutation relation** when computing the Riemann tensor using the tetrad method. Now we discuss its specific expression in the NP formalism. By means of the components $g^{\mu\nu}$ of the metric (inverse) in the null tetrad one can rewrite (5.7.4) as

$$-\gamma^\sigma_{\mu\nu} = g^{\sigma\beta}\omega_{\mu\beta\nu}, \quad (8.7.20)$$

and hence (8.7.19) in the null tetrad becomes

$$[\varepsilon_\mu, \varepsilon_\nu]^a = g^{\sigma\beta}(\omega_{\mu\beta\nu} - \omega_{\nu\beta\mu})(\varepsilon_\sigma)^a. \quad (8.7.21)$$

Take the $\mu\nu$ to be 34, 14, 13, 21, respectively, then the equation above turns into the following 4 commutation relations specifically when applied to a real function (if one also takes $\mu\nu$ to be 24 and 23, the results will be the complex conjugates of the results when taking 14 and 13, and thus are not independent):

$$\Delta D - D\Delta = (\gamma + \bar{\gamma})D + (\varepsilon + \bar{\varepsilon})\Delta - (\tau + \bar{\pi})\bar{\delta} - (\bar{\tau} + \pi)\delta, \quad (8.7.22a)$$

$$\delta D - D\delta = (\bar{\alpha} + \beta - \bar{\pi})D + \kappa\Delta - \sigma\bar{\delta} - (\bar{\rho} + \varepsilon - \bar{\varepsilon})\delta, \quad (8.7.22b)$$

$$\delta\Delta - \Delta\delta = -\bar{v}D + (\tau - \bar{\alpha} - \beta)\Delta + \bar{\lambda}\bar{\delta} + (\mu - \gamma + \bar{\gamma})\delta, \quad (8.7.22c)$$

$$\bar{\delta}\delta - \delta\bar{\delta} = (\bar{\mu} - \mu)D + (\bar{\rho} - \rho)\Delta - (\bar{\alpha} - \beta)\bar{\delta} + (\bar{\beta} - \alpha)\delta. \quad (8.7.22d)$$

When acting on a real function f , (8.7.22a) gives a real equation, (8.7.22d) gives an imaginary equation, and each of (8.7.22b) and (8.7.22c) gives a complex equation; hence, (8.7.22) is equivalent to 6 real equations. To check (8.7.22a), one only has to show that both sides of it acting on any (complex) scalar field f give the same scalar field. It follows from (8.7.9) that

$$\begin{aligned} (\Delta D - D\Delta)f &= (l^b\nabla_b k^a - k^b\nabla_b l^a)\nabla_a f = [l, k]^a\nabla_a f \\ &= [\varepsilon_3, \varepsilon_4]^a\nabla_a f = g^{\sigma\beta}(\omega_{3\beta 4} - \omega_{4\beta 3})(\varepsilon_\sigma)^a\nabla_a f \\ &= [g^{12}(\omega_{324} - \omega_{423})(\varepsilon_1)^a + g^{21}(\omega_{314} - \omega_{413})(\varepsilon_2)^a \\ &\quad + g^{34}(\omega_{344} - \omega_{443})(\varepsilon_3)^a + g^{43}(\omega_{334} - \omega_{433})(\varepsilon_4)^a]\nabla_a f \\ &= \{(-\pi - \bar{\tau})m^a + (-\bar{\pi} - \tau)\bar{m}^a \\ &\quad - [-(\varepsilon + \bar{\varepsilon}) - 0]l^a - [0 - (\gamma + \bar{\gamma})]k^a\}\nabla_a f \\ &= (\gamma + \bar{\gamma})Df + (\varepsilon + \bar{\varepsilon})\Delta f - (\tau + \bar{\pi})\bar{\delta}f - (\bar{\tau} + \pi)\delta f, \end{aligned}$$

and hence we obtain (8.7.22a). The other 3 equations can be verified in a similar manner.

In order to help the readers to better understand the method of solving Einstein's equation using the NP formalism, this text will provide two specific examples in Sect. 8.8.2 and Optional Reading 8.9.1.

8.8 Solving the Einstein-Maxwell Equations Using the NP Formalism [Optional Reading]

8.8.1 Maxwell's Equations and Einstein's Equation in the NP Formalism

Due to the antisymmetry, the electromagnetic tensor F_{ab} has at most 6 independent complex components in the null tetrad, which may be chosen as $F_{43}, F_{42}, F_{41}, F_{32}, F_{31}, F_{21}$. Moreover, they also satisfy the following relations:

$$F_{43} = \bar{F}_{43}, \quad F_{42} = \bar{F}_{41}, \quad F_{32} = \bar{F}_{31}, \quad F_{21} = -F_{12} = -\bar{F}_{21},$$

and thus among all 6 of them, F_{43} and F_{21} are respectively real and imaginary (their sum is complex), and the other 4 are equivalent to two independent complex quantities (we may take F_{41} and F_{23}). Therefore, they are represented by 3 complex quantities Φ_0 , Φ_1 and Φ_2 , defined as

$$\Phi_0 := F_{41} = F_{ab}k^a m^b , \quad (8.8.1a)$$

$$\Phi_1 := \frac{1}{2}(F_{43} + F_{21}) = \frac{1}{2}F_{ab}(k^a l^b + \bar{m}^a m^b) , \quad (8.8.1b)$$

$$\Phi_2 := F_{23} = F_{ab}\bar{m}^a l^b . \quad (8.8.1c)$$

The source-free Maxwell equations

$$\nabla^a F_{ab} = 0 , \quad (8.8.2a)$$

$$\nabla_{[a} F_{bc]} = 0 \quad (8.8.2b)$$

have the following form in the NP formalism:

$$D\Phi_1 - \bar{\delta}\Phi_0 = (\pi - 2\alpha)\Phi_0 + 2\rho\Phi_1 - \kappa\Phi_2 , \quad (8.8.3a)$$

$$D\Phi_2 - \bar{\delta}\Phi_1 = -\lambda\Phi_0 + 2\pi\Phi_1 + (\rho - 2\varepsilon)\Phi_2 , \quad (8.8.3b)$$

$$\delta\Phi_1 - \Delta\Phi_0 = (\mu - 2\gamma)\Phi_0 + 2\tau\Phi_1 - \sigma\Phi_2 , \quad (8.8.3c)$$

$$\delta\Phi_2 - \Delta\Phi_1 = -\nu\Phi_0 + 2\mu\Phi_1 + (\tau - 2\beta)\Phi_2 . \quad (8.8.3d)$$

As an example, here we only provide the verification of (8.8.3a) as follows:

$$\begin{aligned} 2D\Phi_1 &= k^c \nabla_c [F_{ab}(k^a l^b + \bar{m}^a m^b)] = F_{ab}k^a k^c \nabla_c l^b + F_{ab}l^b k^c \nabla_c k^a + k^a l^b k^c \nabla_c F_{ab} \\ &\quad + F_{ab}\bar{m}^a k^c \nabla_c m^b + F_{ab}m^b k^c \nabla_c \bar{m}^a + \bar{m}^a m^b k^c \nabla_c F_{ab} . \end{aligned} \quad (8.8.4)$$

The first and second terms on the right-hand side of the equation above are respectively

$$\begin{aligned} F_{ab}k^a k^c \nabla_c l^b &= F_{4v}(\varepsilon^v)_b (\varepsilon_4)^c \nabla_c (\varepsilon_3)^b = F_{4v}g^{v\mu}\omega_{\mu 34} \\ &= F_{41}g^{12}\omega_{234} + F_{42}g^{21}\omega_{134} + F_{43}g^{34}\omega_{434} = \pi\Phi_0 + \bar{\pi}\bar{\Phi}_0 + F_{43}\omega_{344}, \\ F_{ab}l^b k^c \nabla_c k^a &= -\bar{\kappa}\bar{\Phi}_2 - \kappa\Phi_2 + F_{43}\omega_{344} , \end{aligned}$$

and hence the sum of the first and second terms on the right-hand side of (8.8.4) is $\pi\Phi_0 + \bar{\pi}\bar{\Phi}_0 - \kappa\Phi_2 - \bar{\kappa}\bar{\Phi}_2$. Similarly, the sum of the fourth and fifth terms on the right-hand side of (8.8.4) is $-\kappa\Phi_2 + \bar{\kappa}\bar{\Phi}_2 - \bar{\pi}\bar{\Phi}_0 + \pi\Phi_0$, and therefore,

$$2D\Phi_1 = 2(\pi\Phi_0 - \kappa\Phi_2) + k^a l^b k^c \nabla_c F_{ab} + \bar{m}^a m^b k^c \nabla_c F_{ab} .$$

In a similar manner one can also obtain that

$$\bar{\delta}\Phi_0 = 2(\alpha\Phi_0 - \rho\Phi_1) + k^a m^b \bar{m}^c \nabla_c F_{ab} .$$

Hence,

$$D\Phi_1 - \bar{\delta}\Phi_0 = (\pi - 2\alpha)\Phi_0 + 2\rho\Phi_1 - \kappa\Phi_2 + \frac{1}{2}(k^a l^b k^c + \bar{m}^a m^b k^c - 2k^a m^b \bar{m}^c) \nabla_c F_{ab}. \quad (8.8.5)$$

Let $G \equiv (k^a l^b k^c + \bar{m}^a m^b k^c - 2k^a m^b \bar{m}^c) \nabla_c F_{ab}$, then to verify (8.8.3a) one only has to show that $G = 0$. Maxwell's equations is certainly involved in verifying this. From (8.7.3) we can see that

$$g^{ac} = m^a \bar{m}^c + \bar{m}^a m^c - l^a k^c - k^a l^c, \quad (8.8.6)$$

and hence Maxwell's equation $\nabla^a F_{ab} = 0$ can be written as $(m^a \bar{m}^c + \bar{m}^a m^c - l^a k^c - k^a l^c) \nabla_c F_{ab} = 0$. Contracting this with k^b yields

$$\begin{aligned} 0 &= [m^a k^b \bar{m}^c + \bar{m}^a k^b m^c - (l^a k^b k^c + k^a k^b l^c)] \nabla_c F_{ab} \\ &= [m^a k^b \bar{m}^c - (m^a \bar{m}^b k^c + k^a m^b \bar{m}^c) + k^a l^b k^c] \nabla_c F_{ab} \\ &= [-m^b k^a \bar{m}^c - (-m^b \bar{m}^a k^c + k^a m^b \bar{m}^c) + k^a l^b k^c] \nabla_c F_{ab} = G, \end{aligned}$$

where the second equality is because $\nabla_{[c} F_{ab]} = 0$ leads to $\bar{m}^{[a} k^b m^{c]} \nabla_c F_{ab} = 0$ and $l^{[a} k^b k^{c]} \nabla_c F_{ab} = 0$, and the third equality comes from the fact that $F_{ab} = -F_{ba}$. The other equations in (8.8.3) can be verified similarly.

It follows from (7.2.6) that (Exercise 8.11)

$$\begin{aligned} T_{11} &= \frac{1}{2\pi} \Phi_0 \bar{\Phi}_2, & T_{12} = T_{21} &= \frac{1}{2\pi} \Phi_1 \bar{\Phi}_1, & T_{13} = T_{31} &= \frac{1}{2\pi} \bar{\Phi}_2 \Phi_1, \\ T_{14} = T_{41} &= \frac{1}{2\pi} \Phi_0 \bar{\Phi}_1, & T_{22} &= \frac{1}{2\pi} \Phi_2 \bar{\Phi}_0, & T_{23} = T_{32} &= \frac{1}{2\pi} \Phi_2 \bar{\Phi}_1, \\ T_{24} = T_{42} &= \frac{1}{2\pi} \bar{\Phi}_0 \Phi_1, & T_{33} &= \frac{1}{2\pi} \Phi_2 \bar{\Phi}_2, & T_{34} = T_{43} &= \frac{1}{2\pi} \Phi_1 \bar{\Phi}_1, & T_{44} &= \frac{1}{2\pi} \Phi_0 \bar{\Phi}_0. \end{aligned} \quad (8.8.7)$$

Then, from (8.7.11a), (8.7.11b) and the component form of Einstein's equation $R_{\mu\nu} = 8\pi T_{\mu\nu}$ we obtain the following succinct relations between $\Phi_{00}, \dots, \Phi_{22}$ which represent the curvature tensor and Φ_0, Φ_1, Φ_2 which represent the electromagnetic field tensor:

$$\begin{aligned} \Phi_{00} &= 2\Phi_0 \bar{\Phi}_0, & \Phi_{01} &= 2\Phi_0 \bar{\Phi}_1, & \Phi_{02} &= 2\Phi_0 \bar{\Phi}_2, \\ \Phi_{11} &= 2\Phi_1 \bar{\Phi}_1, & \Phi_{12} &= 2\Phi_1 \bar{\Phi}_2, & \Phi_{22} &= 2\Phi_2 \bar{\Phi}_2. \end{aligned} \quad (8.8.8)$$

This is how Einstein's equation in an electrovac spacetime is expressed in the NP formalism, which can be formulated into the following algebraic equations:

$$\Phi_{\lambda\tau} = 2\Phi_\lambda \bar{\Phi}_\tau, \quad \lambda, \tau = 0, 1, 2. \quad (8.8.9)$$

In Sect. 8.4.1 we introduced a complex quantity Σ to define a null electromagnetic field. It is not difficult to show that (Exercise 8.11) $\Sigma_{ab} \Sigma^{ab}$ can be expressed in terms of the electromagnetic field components Φ_0, Φ_1, Φ_2 in a null tetrad as follows:

$$\Sigma_{ab}\Sigma^{ab} = 16(\Phi_0\Phi_2 - \Phi_1^2). \quad (8.8.10)$$

Hence, the null condition for an electromagnetic field can also be expressed equivalently as

$$\Phi_0\Phi_2 - \Phi_1^2 = 0. \quad (8.8.11)$$

8.8.2 An Example of Solving the Einstein-Maxwell Equations Under the Axisymmetric Condition

In this subsection, we will introduce the detailed process of solving the Einstein-Maxwell equations using the Newman-Penrose formalism by a specific example [see Liang (1995)]. Suppose the metric to be found has the following line element expression in a coordinate system $\{t, z, \varphi, \rho\}$:

$$ds^2 = e^\xi(-dt^2 + d\rho^2) + e^\eta dz^2 + e^{\eta+\chi} d\varphi^2, \quad (8.8.12)$$

where ξ , η and χ are undetermined functions of t and ρ which are independent of z and φ . One can readily see from the equation above that $(\partial/\partial z)^a$ and $(\partial/\partial\varphi)^a$ are two commuting Killing vector fields. Suppose the integral curves of $(\partial/\partial\varphi)^a$ are closed, then (8.8.12) represents a cylindrically symmetric metric, see Sect. 8.5.

Let $v = t + \rho$, $u = t - \rho$, then (8.8.12) becomes

$$ds^2 = -e^\xi du dv + e^\eta dz^2 + e^{\eta+\chi} d\varphi^2, \quad (8.8.13)$$

where ξ , η and χ should be regarded as functions of the new coordinates u and v . Normalizing the orthogonal coordinate basis fields

$$\{(\partial/\partial t)^a, (\partial/\partial\rho)^a, (\partial/\partial z)^a, (\partial/\partial\varphi)^a\}$$

one obtains the orthonormal tetrad fields

$$\begin{aligned} (e_0)^a &= e^{-\xi/2}(\partial/\partial t)^a, & (e_3)^a &= e^{-\xi/2}(\partial/\partial\rho)^a, \\ (e_1)^a &= e^{-\eta/2}(\partial/\partial z)^a, & (e_2)^a &= e^{-(\eta+\chi)/2}(\partial/\partial\varphi)^a. \end{aligned} \quad (8.8.14)$$

By means of (8.7.1), starting from the above orthonormal tetrad fields one can conveniently construct the following null tetrad fields

$$m^a = \frac{1}{\sqrt{2}}[e^{-\eta/2}(\partial/\partial z)^a - ie^{-(\eta+\chi)/2}(\partial/\partial\varphi)^a], \quad (8.8.15a)$$

$$\bar{m}^a = \frac{1}{\sqrt{2}}[e^{-\eta/2}(\partial/\partial z)^a + ie^{-(\eta+\chi)/2}(\partial/\partial\varphi)^a], \quad (8.8.15b)$$

$$l^a = \frac{1}{\sqrt{2}} e^{-\xi/2} [(\partial/\partial t)^a - (\partial/\partial \rho)^a] = \sqrt{2} e^{-\xi/2} (\partial/\partial u)^a, \quad (8.8.15c)$$

$$k^a = \frac{1}{\sqrt{2}} e^{-\xi/2} [(\partial/\partial t)^a + (\partial/\partial \rho)^a] = \sqrt{2} e^{-\xi/2} (\partial/\partial v)^a. \quad (8.8.15d)$$

After computing all the $\omega_{\rho\mu\nu}$ using (5.7.19) [in which the $(e_\mu)^a$ should be interpreted as $(\varepsilon_\mu)^a$] and (5.7.20) or any other method, one can find all (12) complex spin coefficients from (8.7.8) as follows:

$$\kappa = \tau = v = \pi = \beta = \alpha = 0, \quad (8.8.16a)$$

$$\rho = -\frac{\sqrt{2}}{4} e^{-\xi/2} \left(2 \frac{\partial \eta}{\partial v} + \frac{\partial \chi}{\partial v} \right), \quad (8.8.16b)$$

$$\mu = \frac{\sqrt{2}}{4} e^{-\xi/2} \left(2 \frac{\partial \eta}{\partial u} + \frac{\partial \chi}{\partial u} \right), \quad (8.8.16c)$$

$$\varepsilon = \frac{\sqrt{2}}{4} e^{-\xi/2} \frac{\partial \xi}{\partial v}, \quad (8.8.16d)$$

$$\sigma = \frac{\sqrt{2}}{4} e^{-\xi/2} \frac{\partial \chi}{\partial v}, \quad (8.8.16e)$$

$$\lambda = -\frac{\sqrt{2}}{4} e^{-\xi/2} \frac{\partial \chi}{\partial u}, \quad (8.8.16f)$$

$$\gamma = -\frac{\sqrt{2}}{4} e^{-\xi/2} \frac{\partial \xi}{\partial u}. \quad (8.8.16g)$$

When solving the Einstein-Maxwell equations, we have already assumed that there is only an electromagnetic field but no matter fields (“electrovacuum”). The tracelessness of the energy-momentum tensor T_{ab} of the electromagnetic field leads to the fact that the scalar curvature R vanishes. Noticing (8.8.16a), we can see that the NP equations take the following form:

$$D\rho = \rho(\rho + 2\varepsilon) + \sigma^2 + \Phi_{00}, \quad (8.8.17a)$$

$$D\sigma = 2\sigma(\rho + \varepsilon) + \Psi_0, \quad (8.8.17b)$$

$$0 = \Psi_1 + \Phi_{01}, \quad (8.8.17c)$$

$$0 = \Phi_{10}, \quad (8.8.17d)$$

$$0 = \Psi_1, \quad (8.8.17e)$$

$$D\gamma - \Delta\varepsilon = -4\varepsilon\gamma + \Psi_2 + \Phi_{11}, \quad (8.8.17f)$$

$$D\lambda = \lambda(\rho - 2\varepsilon) + \sigma\mu + \Phi_{20}, \quad (8.8.17g)$$

$$D\mu = \mu(\rho - 2\varepsilon) + \sigma\lambda + \Psi_2, \quad (8.8.17h)$$

$$0 = \Psi_3 + \Phi_{21}, \quad (8.8.17i)$$

$$\Delta\lambda = -2\lambda(\mu + \gamma) - \Psi_4, \quad (8.8.17j)$$

$$0 = -\Psi_1 + \Phi_{01}, \quad (8.8.17k)$$

$$0 = \mu\rho - \lambda\sigma - \Psi_2 + \Phi_{11}, \quad (8.8.17l)$$

$$0 = -\Psi_3 + \Phi_{21}, \quad (8.8.17m)$$

$$-\Delta\mu = \mu(\mu + 2\gamma) + \lambda^2 + \Phi_{22}, \quad (8.8.17n)$$

$$0 = \Phi_{12}, \quad (8.8.17o)$$

$$-\Delta\sigma = \sigma(\mu - 2\gamma) + \lambda\rho + \Phi_{02}, \quad (8.8.17p)$$

$$\Delta\rho = \rho(2\gamma - \mu) - \sigma\lambda - \Psi_2, \quad (8.8.17q)$$

$$0 = -\Psi_3. \quad (8.8.17r)$$

Our discussion is limited only to the case of a source-free electromagnetic field, and hence when (8.8.16a) holds Maxwell's equations will take the following form:

$$D\Phi_1 - \bar{\delta}\Phi_0 = 2\rho\Phi_1, \quad (8.8.18a)$$

$$D\Phi_2 - \bar{\delta}\Phi_1 = -\lambda\Phi_0 + (\rho - 2\varepsilon)\Phi_2, \quad (8.8.18b)$$

$$\delta\Phi_1 - \Delta\Phi_0 = (\mu - 2\gamma)\Phi_0 - \sigma\Phi_2, \quad (8.8.18c)$$

$$\delta\Phi_2 - \Delta\Phi_1 = 2\mu\Phi_1. \quad (8.8.18d)$$

From Einstein's equations (8.8.9) we can see that (8.8.17d) and (8.8.17o) will lead to $\Phi_1 = 0$ or $\Phi_0 = \Phi_2 = 0$. It follows from the null condition $\Phi_0\Phi_2 - \Phi_1^2 = 0$ that an electromagnetic field with $\Phi_0 = \Phi_2 = 0$ can only be a nonnull electromagnetic field, while an electromagnetic field with $\Phi_1 = 0$ can be either null or nonnull. Here we only discuss nonnull electromagnetic fields with $\Phi_1 = 0$; that is, we only seek for the solutions of nonnull electromagnetic fields with $\Phi_1 = 0$ (which must have $\Phi_0 \neq 0$ and $\Phi_2 \neq 0$). In this case, Maxwell's equations (8.8.18) will be simplified to

$$\bar{\delta}\Phi_0 = 0, \quad (8.8.19a)$$

$$D\Phi_2 = -\lambda\Phi_0 + (\rho - 2\varepsilon)\Phi_2, \quad (8.8.19b)$$

$$-\Delta\Phi_0 = (\mu - 2\gamma)\Phi_0 - \sigma\Phi_2, \quad (8.8.19c)$$

$$\delta\Phi_2 = 0. \quad (8.8.19d)$$

By solving the Einstein-Maxwell equations, we mean finding the expression of the metric functions $\xi(t, \rho), \eta(t, \rho), \chi(t, \rho)$ as well as the electromagnetic field functions Φ_0 and Φ_2 which satisfy these equations. They appear in the following 3 systems of equations (which are coupled with each other): ① Maxwell's equations (8.8.19); ② Einstein's equations $\Phi_{\lambda\tau} = 2\Phi_\lambda\dot{\Phi}_\tau$ ($\lambda, \tau = 0, 1, 2$); ③ the NP equations (8.8.17). The solving process is as follows.

Equation (8.8.19d) will lead to

$$\frac{\partial\Phi_2}{\partial z} - ie^{-\chi/2} \frac{\partial\Phi_2}{\partial\varphi} = 0. \quad (8.8.20)$$

However, one cannot yet say that $\partial\Phi_2/\partial z = \partial\Phi_2/\partial\varphi = 0$ since Φ_2 is a complex-valued function. Suppose $\Phi_2 = Ce^{i\theta}$, where C and θ are real-valued functions, then

$$\Phi_{22} = 2\Phi_2\bar{\Phi}_2 = 2C^2. \quad (8.8.21)$$

Since μ, γ, λ are all independent of z and φ , (8.8.17n) indicates that C is independent of z and φ , and hence (8.8.20) gives

$$\left(\frac{\partial}{\partial z} - ie^{-\chi/2} \frac{\partial}{\partial \varphi} \right) e^{i\theta} = 0.$$

Thus,

$$\frac{\partial \theta}{\partial z} = \frac{\partial \theta}{\partial \varphi} = 0,$$

i.e., Φ_2 is indeed independent of z and φ . Similarly, it follows from (8.8.19a), (8.8.17a) and Einstein's equation $\Phi_{00} = 2\Phi_0\bar{\Phi}_0$ that Φ_0 is also independent of z and φ . On the other hand, (8.8.19b) and (8.8.19c) can be expressed as

$$-4 \frac{\partial \Phi_2}{\partial v} = \left(2 \frac{\partial \xi}{\partial v} + 2 \frac{\partial \eta}{\partial v} + \frac{\partial \chi}{\partial v} \right) \Phi_2 - \frac{\partial \chi}{\partial u} \Phi_0, \quad (8.8.22)$$

$$-4 \frac{\partial \Phi_0}{\partial u} = \left(2 \frac{\partial \xi}{\partial u} + 2 \frac{\partial \eta}{\partial u} + \frac{\partial \chi}{\partial u} \right) \Phi_0 - \frac{\partial \chi}{\partial v} \Phi_2. \quad (8.8.23)$$

To make the solving process more tractable, we will only discuss the case where $\partial\chi/\partial u = 0$. As long as we have a solution under this condition, we will obtain an exact solution. Of course, we cannot assure beforehand that there must be a solution in this case, and so this is a tentative approach. Now we only have to care about the case $\partial\chi/\partial v \neq 0$. This is because $\partial\chi/\partial u = \partial\chi/\partial v = 0$ will make the line element (8.8.13) locally the same as a plane symmetric metric, and the plane symmetric metrics generated by “semi-plane symmetric” (which locally looks like cylindrically symmetric) electromagnetic fields have been exhausted by Li and Liang (1985). The condition $\partial\chi/\partial u = 0$ brings us many simplifications, for instance it leads to $\lambda = 0$, also one can now integrate (8.8.22) and get

$$\Phi_2(u, v) = a(u)e^{-(2\xi+2\eta+\chi)/4}, \quad (8.8.24)$$

where $a(u)$ is an arbitrary complex-valued function of u , and $a(u) \neq 0$. (Otherwise $\Phi_2 = 0$, which contradicts the premise). Hence, it follows from Einstein's equation $\Phi_{22} = 2\Phi_2\bar{\Phi}_2$ that

$$\Phi_{22}(u, v) = 2|a(u)|^2 e^{-(2\xi+2\eta+\chi)/2}. \quad (8.8.25)$$

There is now only one unsolved Maxwell equation remaining, namely (8.8.23), which can be simplified as

$$-4 \frac{\partial \Phi_0}{\partial u} = 2 \left(\frac{\partial \xi}{\partial u} + \frac{\partial \eta}{\partial u} \right) \Phi_0 - \chi' a(u) e^{-(2\xi+2\eta+\chi)/4}, \quad (8.8.26)$$

where the ' represents the derivative of a function of one variable (for the above equation it is $\chi' \equiv d\chi/dv$). The condition $\partial \chi / \partial u = 0$ also simplifies the NP equations, for instance (8.8.17g) now becomes

$$-\Phi_{20} = \frac{1}{4} e^{-\xi} \chi' \frac{\partial \eta}{\partial u}, \quad (8.8.27)$$

which says that Φ_{20} is a real number, and thus $\Phi_{02} = \Phi_{20}$. Noticing that $a(u) \neq 0$ (otherwise the electromagnetic field vanishes), by combining (8.8.27) with (8.8.9) and (8.8.24) we get

$$\Phi_0(u, v) = -\frac{1}{8\bar{a}(u)} \chi' \frac{\partial \eta}{\partial u} e^{(-2\xi+2\eta+\chi)/4}. \quad (8.8.28)$$

Taking the derivative of the above equation with respect to u and plugging into (8.8.26) we obtain

$$-2|a|^2 = \left[-\bar{a}^{-1} \bar{a}' \frac{\partial \eta}{\partial u} + \frac{\partial^2 \eta}{\partial u^2} + \left(\frac{\partial \eta}{\partial u} \right)^2 \right] e^{\eta+\chi/2}. \quad (8.8.29)$$

Now we look back at the NP equations (8.8.17). Equation (g) has been used. By means of (8.8.16) and (8.8.27) it is not difficult to verify that (p) is automatically satisfied. The assumption that $\Phi_1 = 0$ leads to $\Phi_{01} = \Phi_{10} = \Phi_{12} = \Phi_{21} = \Phi_{11} = 0$, and so (d) and (o) become identities; also, (c) becomes equivalent to (k) and (e), which states nothing but the fact that the Weyl tensor of the spacetime has its component

$$\Psi_1 = 0. \quad (8.8.30)$$

Similarly, (i), (m) and (r) being equivalent gives

$$\Psi_3 = 0. \quad (8.8.31)$$

In addition, $\lambda = 0$ simplifies (j) and (l) a lot and gives

$$\Psi_4 = 0, \quad (8.8.32)$$

$$\Psi_2 = \mu\rho. \quad (8.8.33)$$

If we leave (l) [i.e., (8.8.33)] and (b) to the end to determine Ψ_2 and Ψ_0 (no need to solve), then the NP equations (8.8.17) has only 5 unsolved equations remaining, namely (a), (f), (h), (n) and (q). Noticing $\Phi_{11} = 0$, $\lambda = 0$ and (8.8.33), we see that these 5 equations take the following form:

$$D\rho = \rho(\rho + 2\varepsilon) + \sigma^2 + \Phi_{00}, \quad (8.8.34)$$

$$D\gamma - \Delta\varepsilon = -4\varepsilon\gamma + \mu\rho, \quad (8.8.35)$$

$$D\mu = 2\mu(\rho - \varepsilon), \quad (8.8.36)$$

$$-\Delta\mu = \mu(\mu + 2\gamma) + \Phi_{22}, \quad (8.8.37)$$

$$\Delta\rho = 2\rho(\gamma - \mu). \quad (8.8.38)$$

Equations (8.8.36) and (8.8.38) are both equivalent to

$$\frac{\partial^2\eta}{\partial u\partial v} = -\frac{\partial\eta}{\partial u}\left(\frac{\partial\eta}{\partial v} + \frac{1}{2}\chi'\right),$$

integrating this yields

$$\eta(u, v) = -\frac{1}{2}\chi + \ln[g(v) - f(u)], \quad (8.8.39)$$

where $g(v)$ and $f(u)$ are arbitrary functions. Hence, (8.8.35) becomes

$$\frac{\partial^2\xi}{\partial u\partial v} = -\frac{1}{2}(g - f)^{-2}f'g',$$

integrating this yields

$$\xi(u, v) = -\frac{1}{2}\ln(g - f) + F(u) + G(v), \quad (8.8.40)$$

where $F(u)$ and $G(v)$ are arbitrary functions. Plugging (8.8.39) and (8.8.40) into (8.8.13) yields

$$ds^2 = -(g - f)^{-1/2}e^{F+G}dudv + (g - f)(e^{-\chi/2}dz^2 + e^{\chi/2}d\varphi^2). \quad (8.8.41)$$

Define new coordinates \tilde{u} and \tilde{v} as follows: $d\tilde{u} = e^{F(u)}du$, $d\tilde{v} = e^{G(v)}dv$, then

$$ds^2 = -(g - f)^{-1/2}d\tilde{u}d\tilde{v} + (g - f)(e^{-\chi/2}dz^2 + e^{\chi/2}d\varphi^2). \quad (8.8.42)$$

If we take $F(u) = G(v) = 0$, then it follows from (8.8.41) that

$$ds^2 = -(g - f)^{-1/2}dudv + (g - f)(e^{-\chi/2}dz^2 + e^{\chi/2}d\varphi^2). \quad (8.8.42')$$

Equations (8.8.42) and (8.8.42') represent the same line element (the only difference is the coordinate notations u and v are changed to \tilde{u} and \tilde{v} , which is not essential), and thus when taking $F(u) = G(v) = 0$ we do not lose any solution. Henceforth we will take this choice, i.e., take (8.8.42') as the line element.

Now, 3 unsolved equations remain, namely (8.8.29), (8.8.34) and (8.8.37), and the undetermined functions are $g(v)$, $f(u)$, $\chi(v)$ and $a(u)$. Equation (8.8.37) is equivalent to

$$\frac{\partial^2 \eta}{\partial u^2} - \frac{\partial \xi}{\partial u} \frac{\partial \eta}{\partial u} + \frac{1}{2} \left(\frac{\partial \eta}{\partial u} \right)^2 + 2|a|^2 e^{-(\eta+\chi/2)} = 0. \quad (8.8.43)$$

By means of (8.8.39) and (8.8.40) (where $F = G = 0$) one can rewrite the equation above as

$$f'' = 2|a(u)|^2. \quad (8.8.44)$$

Plugging (8.8.38) into (8.8.29) yields

$$\bar{a}^{-1} \bar{a}' f' = f'' - 2|a|^2 = 0, \quad (8.8.45)$$

where (8.8.44) is used in the second equality. The equation above indicates that either $a' = 0$ or $f' = 0$; however, from (8.8.44) we know that the latter leads to $a = 0$, which is not allowed, and hence we have only $a' = 0$, i.e., $a = \text{constant}$. Thus, integrating (8.8.44) yields

$$f' = 2A^2 u + c_1, \quad f = A^2 u^2 + c_1 u + c_2, \quad (8.8.46)$$

where $A \equiv |a|$, and c_1, c_2 are real constants of integration.

Now let us consider the last unsolved Maxwell equation, namely (8.8.34). It follows from (8.8.28), (8.8.29), (8.8.40) and (8.8.9) that

$$\Phi_{00} = (32|a|^2)^{-1} (g - f)^{-1/2} \chi'^2 f'^2. \quad (8.8.47)$$

Plugging this into (8.8.34), by a brief calculation we can see that (8.8.34) is equivalent to

$$8g''(v)\chi'^{-2}(v) + g(v) = f(u) - (4|a|^2)^{-1} f'^2(u). \quad (8.8.48)$$

In the above equation, the left-hand side is not a function of u and the right-hand side is not a function of v , and thus both sides are equal to a constant, denoted by K , i.e.,

$$8g''(v)\chi'^{-2}(v) + g(v) = K, \quad (8.8.49)$$

$$f(u) - (4A^2)^{-1} f'^2(u) = K. \quad (8.8.50)$$

Plugging (8.8.46) into (8.8.50) yields

$$K = c_2 - (4A^2)^{-1} c_1^2. \quad (8.8.51)$$

Therefore, the line element (8.8.42') can be expressed as

$$\begin{aligned} ds^2 = & -[g(v) - A^2 u^2 - c_1 u - c_2]^{-1/2} du dv \\ & + [g(v) - A^2 u^2 - c_1 u - c_2] (e^{-\chi(v)/2} dz^2 + e^{\chi(v)/2} d\varphi^2), \end{aligned} \quad (8.8.52)$$

where A , c_1 and c_2 are arbitrary constants, the functions $g(v)$ and $\chi(v)$ are quite arbitrary but are related by (8.8.48), in which the value K on both sides depends on our choice of the constants A , c_1 and c_2 .

Conclusion: After choosing the constants A , c_1 and c_2 , any real function pair $(g(v), \chi(v))$ satisfying (8.8.49) determines a cylindrically symmetric metric by (8.8.52), whose corresponding source is a cylindrically symmetric nonnull electromagnetic field described by a complex-valued function pair (Φ_0, Φ_2) satisfying (8.8.28) and (8.8.24). There are many real function pairs $(g(v), \chi(v))$ that satisfy (8.8.49), for instance the following 3 function pairs all satisfy (8.8.49) with c_1 and c_2 chosen to be zero, i.e., $K = 0$:

$$(1) g(v) = \sin v, \chi(v) = 2\sqrt{2}v.$$

$$(2) g(v) = \ln v, \chi(v) = 4\sqrt{2}(\ln v)^{1/2}.$$

(3) $g(v) = v^{1/\alpha}$, $\chi(v) = (2/\alpha)\sqrt{2(\alpha-1)} \ln v$, where $\alpha \in (1, \infty)$. This example forms a one-parameter subfamily (with the parameter α) of the cylindrically symmetric solution family of the Einstein-Maxwell equations, in which the simplest one is the solution characterized by $\alpha = 2$, i.e., $g(v) = v^{1/2}$, $\chi(v) = \sqrt{2} \ln v$.

The electromagnetic field F_{ab} described by a complex-valued function pair (Φ_0, Φ_2) satisfying (8.8.28) and (8.8.24) can also be expressed in terms of its non-vanishing components in the coordinate basis $\{(\partial/\partial t)^a, (\partial/\partial\rho)^a, (\partial/\partial z)^a, (\partial/\partial\varphi)^a\}$:

$$F_{tz} = -F_{zt} = -a_1 e^{-\chi/4} \left(1 - \frac{1}{4} u \chi' \right), \quad (8.8.53a)$$

$$F_{\rho z} = -F_{z\rho} = a_1 e^{-\chi/4} \left(1 + \frac{1}{4} u \chi' \right), \quad (8.8.53b)$$

$$F_{t\varphi} = -F_{\varphi t} = -a_2 e^{\chi/4} \left(1 + \frac{1}{4} u \chi' \right), \quad (8.8.53c)$$

$$F_{\rho\varphi} = -F_{\varphi\rho} = a_2 e^{\chi/4} \left(1 - \frac{1}{4} u \chi' \right), \quad (8.8.53d)$$

where $F_{tz} \equiv F_{ab}(\partial/\partial t)^a(\partial/\partial z)^b$, the others are defined similarly; $a_1, a_2 \in \mathbb{R}$ are the real and imaginary parts of a , respectively. It is not difficult to verify that F_{ab} constituted by (8.8.53) satisfies the source-free Maxwell equations $\nabla^a F_{ab} = 0$ and $\nabla_{[a} F_{bc]} = 0$, and the energy-momentum tensor T_{ab} constituted by F_{ab} according to (8.4.1) satisfies Einstein's equation $T_{ab} = R_{ab}/8\pi$, where R_{ab} is the Ricci tensor of the metric (8.8.52).

8.9 The Vaidya Metric and the Kinnersley Metric

8.9.1 From the Schwarzschild Metric to the Vaidya Metric

The line element of the vacuum Schwarzschild solution in the Schwarzschild coordinate system $\{t, r, \theta, \varphi\}$ is given by

$$ds_{\text{Sch}}^2 = -\left(1 - \frac{2M}{r}\right)dt^2 + \left(1 - \frac{2M}{r}\right)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2) \quad (r > 2M).$$

Starting from the Schwarzschild coordinate system, we apply the coordinate transformation $\{t, r, \theta, \varphi\} \mapsto \{u, r, \theta, \varphi\}$, where

$$u \equiv t - r_*, \quad r_* \equiv r + 2M \ln\left(\frac{r}{2M} - 1\right) \quad (r_* \text{ is called the } \mathbf{\text{tortoise coordinate}}), \quad (8.9.1)$$

then the Schwarzschild line element turns into the following form:

$$\begin{aligned} ds_{\text{Sch}}^2 &= -(1 - 2Mr^{-1})du^2 - 2dudr + r^2(d\theta^2 + \sin^2\theta d\varphi^2) \\ &= [-du^2 - 2dudr + r^2(d\theta^2 + \sin^2\theta d\varphi^2)] + 2Mr^{-1}du^2. \end{aligned} \quad (8.9.2)$$

The square bracket on the right-hand side of the above equation can also be written as $-dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2)$, which is nothing but the flat line element, and hence

$$ds_{\text{Sch}}^2 = ds_{\text{flat}}^2 + 2Mr^{-1}du^2. \quad (8.9.3)$$

Once we change the constant M in the above equation to a function $m(u)$ of the coordinate u , we obtain the following new line element [called the **Vaidya line element**]:

$$\begin{aligned} ds_{\text{Vai}}^2 &= ds_{\text{flat}}^2 + 2m(u)r^{-1}du^2 \\ &= -[1 - 2m(u)r^{-1}]du^2 - 2dudr + r^2(d\theta^2 + \sin^2\theta d\varphi^2). \end{aligned} \quad (8.9.4)$$

Let g_{ab} represent the Vaidya metric, then from the equation above one can read off all of its nonvanishing components in the system $\{u, r, \theta, \varphi\}$:

$$g_{uu} = -[1 - 2m(u)r^{-1}], \quad g_{ur} = g_{ru} = -1, \quad g_{\theta\theta} = r^2, \quad g_{\varphi\varphi} = r^2 \sin^2\theta. \quad (8.9.5)$$

Hence, from $g_{ab} = g_{\mu\nu}(dx^\mu)_a(dx^\nu)_b$ we get the abstract index expression for the Vaidya metric:

$$\begin{aligned} g_{ab} &= -[1 - 2m(u)r^{-1}](du)_a(du)_b - (du)_a(dr)_b - (dr)_a(du)_b \\ &\quad + r^2(d\theta)_a(d\theta)_b + r^2 \sin^2\theta(d\varphi)_a(d\varphi)_b. \end{aligned} \quad (8.9.6)$$

It is not difficult to verify that the inverse of g_{ab} is

$$\begin{aligned} g^{ab} = & -\left(\frac{\partial}{\partial u}\right)^a \left(\frac{\partial}{\partial r}\right)^b - \left(\frac{\partial}{\partial r}\right)^a \left(\frac{\partial}{\partial u}\right)^b + \left[1 - \frac{2m(u)}{r}\right] \left(\frac{\partial}{\partial r}\right)^a \left(\frac{\partial}{\partial r}\right)^b \\ & + \frac{1}{r^2} \left(\frac{\partial}{\partial \theta}\right)^a \left(\frac{\partial}{\partial \theta}\right)^b + \frac{1}{r^2 \sin^2 \theta} \left(\frac{\partial}{\partial \varphi}\right)^a \left(\frac{\partial}{\partial \varphi}\right)^b. \end{aligned} \quad (8.9.7)$$

Now that we have the metric we can compute its Einstein tensor $G_{ab} \equiv R_{ab} - Rg_{ab}/2$, and from Einstein's equation $G_{ab} = 8\pi T_{ab}$ we can find its energy-momentum tensor T_{ab} in order to figure out what is the source associated with this metric. The nonvanishing components of the Vaidya metric are already given in (8.9.5), and the corresponding inverse matrix has the nonvanishing components

$$g^{rr} = 1 - 2m(u)r^{-1}, \quad g^{ur} = g^{ru} = -1, \quad g^{\theta\theta} = r^{-2}, \quad g^{\varphi\varphi} = (r \sin \theta)^{-2}. \quad (8.9.8)$$

Plugging them into (3.2.10') yields the nonvanishing Christoffel symbols

$$\begin{aligned} \Gamma^u_{uu} &= -mr^{-2}, \quad \Gamma^u_{\theta\theta} = r, \quad \Gamma^u_{\varphi\varphi} = r \sin^2 \theta \\ \Gamma^r_{uu} &= -\dot{m}r^{-1} + mr^{-3}(r - 2m), \quad \Gamma^r_{ur} = \Gamma^r_{ru} = mr^{-2}, \\ \Gamma^r_{\theta\theta} &= 2m - r, \quad \Gamma^r_{\varphi\varphi} = (2m - r) \sin^2 \theta, \\ \Gamma^\theta_{r\theta} &= \Gamma^\theta_{\theta r} = r^{-1}, \quad \Gamma^\theta_{\varphi\varphi} = -\sin \theta \cos \theta, \\ \Gamma^\varphi_{r\varphi} &= \Gamma^\varphi_{\varphi r} = r^{-1}, \quad \Gamma^\varphi_{\theta\varphi} = \Gamma^\varphi_{\varphi\theta} = \cot \theta, \end{aligned} \quad (8.9.9)$$

where $\dot{m} \equiv dm(u)/du$. Plugging these into (3.4.21), we see that the Ricci tensor R_{ab} has only one nonvanishing component, i.e.,

$$R_{uu} = -2\dot{m}r^{-2}, \quad (8.9.10)$$

and hence

$$R_{ab} = -2\dot{m}r^{-2}(\mathrm{d}u)_a(\mathrm{d}u)_b. \quad (8.9.11)$$

From the above equation we also get $R = g^{uu}R_{uu} = 0$, and hence $G_{ab} = R_{ab}$. Therefore, it follows from Einstein's equation $G_{ab} = 8\pi T_{ab}$ that

$$T_{ab} = -\frac{\dot{m}}{4\pi r^2}(\mathrm{d}u)_a(\mathrm{d}u)_b. \quad (8.9.12)$$

Let

$$k_a \equiv -(\mathrm{d}u)_a, \quad k^a \equiv g^{ab}k_b = -g^{ab}(\mathrm{d}u)_b, \quad (8.9.13)$$

then from (8.9.7) we can easily see that

$$k^a = (\partial/\partial r)^a. \quad (8.9.14)$$

Hence, $k^a k_a = 0$, and thus k^a is a null vector field. Now (8.9.12) can also be expressed as

$$T_{ab} = -\frac{\dot{m}}{4\pi r^2} k_a k_b. \quad (8.9.12')$$

When $\dot{m} < 0$, the above equation can be viewed as a special case of the energy-momentum tensor in the following form:

$$T_{ab} = \Phi^2 k_a k_b \quad (\Phi^2 \text{ is a positive definite function}). \quad (8.9.15)$$

What kind of field can have an energy-momentum tensor like that shown in the above equation? It can be proved that (see Appendix D in Volume II) the energy-momentum tensor of a source-free null electromagnetic field (satisfying $F_{ab} F^{ab} = 0$) can be expressed in the form (8.9.15), in which

$$\Phi^2 \equiv \frac{E^2}{2\pi} \quad (E \text{ is the electric field measured by an orthonormal tetrad}). \quad (8.9.16)$$

A null electromagnetic field can be viewed as a “matter field” formed by many photons propagating along the null direction k^a . Moreover, a matter field formed by other particles with zero rest masses (such as massless scalar particles and neutrinos¹⁰) moving along the k^a -direction also has an energy-momentum tensor of the form (8.9.15). This kind of matter field is called a **pure radiation field**. In summary, the matter fields whose energy-momentum tensors can be expressed in terms of (8.9.15) can be classified into two kinds: ① source-free null electromagnetic fields; ② pure radiation fields. The difference between them is that there exists a 2-form field F_{ab} for the former one, which satisfies the source-free Maxwell equations and $T_{ab} = F_{ac} F_b{}^c / 4\pi$. It can be proved that (see Optional Reading 8.9.1) the matter field corresponding to (8.9.12') does not obey the source-free Maxwell equations, and thus the source of the Vaidya metric is a pure radiation field instead of a null electromagnetic field.

When compared with the Schwarzschild metric, the Vaidya metric has mainly the following three differences. ① The mass parameter M of the former one is a constant while the m in the latter is a function of u . ② The former is a solution to the vacuum Einstein equation $G_{ab} = 0$ while the latter is a solution of the Einstein equation with source $G_{ab} = 8\pi T_{ab}$, where T_{ab} represents a pure radiation field. ③ By finding the general solution to the Killing equation one can show that, the former has four independent Killing vector fields, in which one of them is timelike, and hence is a stationary metric; the latter has only three independent Killing vector fields (which are exactly those three reflecting spherical symmetry) with no timelike Killing field, and hence the Vaidya solution is not a stationary metric. The above three properties of the Vaidya metric are closely related. If we interpret m still as the

¹⁰ This is included because neutrinos are massless in the Standard Model of particle physics. However, now it has been experimentally confirmed that neutrinos have nonzero masses, and thus technically it should not be included anymore.

mass of a spherically symmetric star, and interpret u as the proper time of the star (Sect. 8.9.3 will justify this interpretation), then m being a function of u (property ①) indicates that the mass of the star changes with time with a rate \dot{m} . Why is it so? Because it keeps emitting massless particles (property ②) (for convenience they are also called “photons”, although they are not quanta of an electromagnetic field), which takes away energy ceaselessly. Calculation (see Sect. 8.9.3) shows that the energy flows to infinity per unit time happen to be equal to $-\dot{m}$, i.e., equal to the decreasing rate of the energy (mass) m of the star (assuming $\dot{m} < 0$),¹¹ which agrees with the law of the conservation of energy. It is exactly the feature that m is time dependent which renders the Vaidya metric a non-stationary metric (property ③). In consideration of the above-mentioned properties, P. C. Vaidya himself called this kind of star a “shining star”, although the “shining” is not caused by photons but other massless particles. It is natural to ask: does not a static star described by the Schwarzschild metric shine? Of course a star shines, but the thing is, to simplify the solving process, Schwarzschild ignored the energy-momentum tensor of the photons emitted from the star (which also form a bath for the star) and treated its exterior as a vacuum. This is how we can have the well-known, exceptionally easy, while extensively used, vacuum Schwarzschild solution. Thus, the familiar physical interpretation “the vacuum Schwarzschild solution describes the exterior metric field of a static spherically symmetric star” is only an approximate statement.

[Optional Reading 8.9.1]

As another example of applying the NP formalism, here we compute the Riemann tensor of the Vaidya metric again using the null tetrad. The first step is to choose an appropriate null tetrad $\{\varepsilon_\mu\}^a$. The expression (8.9.6) for the Vaidya metric g_{ab} can also be written as

$$\begin{aligned} g_{ab} = & -h(\mathrm{d}u)_a(\mathrm{d}u)_b - (\mathrm{d}u)_a(\mathrm{d}r)_b - (\mathrm{d}r)_a(\mathrm{d}u)_b \\ & + r^2(\mathrm{d}\theta)_a(\mathrm{d}\theta)_b + r^2 \sin^2 \theta(\mathrm{d}\varphi)_a(\mathrm{d}\varphi)_b, \end{aligned} \quad (8.9.17)$$

where

$$h \equiv 1 - \frac{2m(u)}{r}. \quad (8.9.18)$$

The following reformulation of the above expression will bring us important inspiration:

$$\begin{aligned} g_{ab} = & -(\mathrm{d}u)_a \left[\frac{1}{2}h(\mathrm{d}u)_b + (\mathrm{d}r)_b \right] - \left[\frac{1}{2}h(\mathrm{d}u)_a + (\mathrm{d}r)_a \right] (\mathrm{d}u)_b \\ & + \{r[(\mathrm{d}\theta)_a - i \sin \theta(\mathrm{d}\varphi)_a]/\sqrt{2}\}\{r[(\mathrm{d}\theta)_b + i \sin \theta(\mathrm{d}\varphi)_b]/\sqrt{2}\} \\ & + \{r[(\mathrm{d}\theta)_a + i \sin \theta(\mathrm{d}\varphi)_a]/\sqrt{2}\}\{r[(\mathrm{d}\theta)_b - i \sin \theta(\mathrm{d}\varphi)_b]/\sqrt{2}\}. \end{aligned} \quad (8.9.19)$$

Comparing with the general expression in the null tetrad $\{\varepsilon_\mu\}^a\}$

$$g_{ab} = g_{\mu\nu}(\varepsilon^\mu)_a(\varepsilon^\nu)_b = -k_a l_b - l_a k_b + \bar{m}_a m_b + m_a \bar{m}_b, \quad (8.9.20)$$

we can “read off” m_a, \bar{m}_a, l_a and k_a as follows:

¹¹ As a solution to Einstein’s equation, the derivative of the parameter $m(u)$ can either be positive or negative (also, of course, zero). However, in order to make this solution a metric corresponding to a matter field which is physically acceptable, we need to require $\dot{m} < 0$.

$$\begin{aligned} k_a &= -(\text{d}u)_a, \quad l_a = -\frac{1}{2}h(\text{d}u)_a - (\text{d}r)_a, \\ m_a &= \frac{r}{\sqrt{2}}[(\text{d}\theta)_a - i \sin \theta (\text{d}\varphi)_a], \quad \bar{m}_a = \frac{r}{\sqrt{2}}[(\text{d}\theta)_a + i \sin \theta (\text{d}\varphi)_a], \end{aligned} \quad (8.9.21)$$

and the corresponding m^a, \bar{m}^a, l^a and k^a are

$$\begin{aligned} k^a &= (\partial/\partial r)^a, \quad m^a = \frac{1}{\sqrt{2}r}[(\partial/\partial\theta)^a - i \sin^{-1} \theta (\partial/\partial\varphi)^a], \\ l^a &= (\partial/\partial u)^a - \frac{1}{2}h(\partial/\partial r)^a, \quad \bar{m}^a = \frac{1}{\sqrt{2}r}[(\partial/\partial\theta)^a + i \sin^{-1} \theta (\partial/\partial\varphi)^a]. \end{aligned} \quad (8.9.21')$$

The readers should verify that this null tetrad indeed satisfies

$$gabm^a m^b = gab\bar{m}^a \bar{m}^b = gabl_a l^b = gabk^a k^b = 0, \quad gabm^a \bar{m}^b = 1, \quad gabl^a k^b = -1.$$

After computing all of the $\omega_{\rho\mu\nu}$ using (5.7.19) [in which $(e_\mu)^a$ should be interpreted as $(\varepsilon_\mu)^a$] and (5.7.20) or any other method, one can find all of the 12 spin coefficients using (8.7.8) as follows:

$$\kappa = \sigma = \nu = \tau = \lambda = \pi = \varepsilon = 0, \quad (8.9.22a)$$

$$\rho = -\frac{1}{r}, \quad \mu = -\frac{1}{2r} \left[1 - \frac{2m(u)}{r} \right], \quad \gamma = \frac{m}{2r^2}, \quad \beta = -\alpha = \frac{1}{2\sqrt{2}r} \cot \theta. \quad (8.9.22b)$$

Using (8.9.22a) one can simplify the NP equations into the following form:

$$D\rho = \rho^2 + \Phi_{00}, \quad (8.9.23a)$$

$$0 = \Psi_0, \quad (8.9.23b)$$

$$0 = \Psi_1 + \Phi_{01}, \quad (8.9.23c)$$

$$D\alpha = \alpha\rho + \Phi_{10}, \quad (8.9.23d)$$

$$D\beta = \beta\bar{\rho} + \Psi_1, \quad (8.9.23e)$$

$$D\gamma = \Psi_2 + \Phi_{11} - R/24, \quad (8.9.23f)$$

$$0 = \Phi_{20}, \quad (8.9.23g)$$

$$D\mu = \bar{\rho}\mu + \Psi_2 + R/12, \quad (8.9.23h)$$

$$0 = \psi_3 + \Phi_{21}, \quad (8.9.23i)$$

$$0 = -\Psi_4, \quad (8.9.23j)$$

$$\delta\rho = \rho(\bar{\alpha} + \beta) - \Psi_1 + \Phi_{01}, \quad (8.9.23k)$$

$$\delta\alpha - \bar{\delta}\beta = \mu\rho + (\alpha\bar{\alpha} + \beta\bar{\beta} - 2\alpha\beta) - \Psi_2 + \Phi_{11} + R/24, \quad (8.9.23l)$$

$$-\bar{\delta}\mu = -\Psi_3 + \Phi_{21}, \quad (8.9.23m)$$

$$-\Delta\mu = \mu^2 + \mu(\gamma + \bar{\gamma}) + \Phi_{22}, \quad (8.9.23n)$$

$$-\Delta\beta = \gamma(-\bar{\alpha} - \beta) - \beta(\gamma - \bar{\gamma} - \mu) + \Phi_{12}, \quad (8.9.23o)$$

$$0 = \Phi_{02}, \quad (8.9.23p)$$

$$\Delta\rho = -\rho\bar{\mu} + \rho(\gamma + \bar{\gamma}) - \Psi_2 - R/12, \quad (8.9.23q)$$

$$\Delta\alpha = \alpha(\bar{\gamma} - \bar{\mu}) + \gamma\bar{\beta} - \Psi_3. \quad (8.9.23r)$$

Plugging (8.9.22b) into (8.9.23), one can readily find the 5 complex quantities $\Psi_0 \sim \Psi_4$ representing the Weyl tensor and the 4 real quantities $\Phi_{00}, \Phi_{11}, \Phi_{22}, R$ representing the Ricci tensor as well as 3 independent complex quantities $\Phi_{01}, \Phi_{02}, \Phi_{12}$. Among them only two are nonvanishing:

$$\Psi_2 = -m(u)/r^3, \quad (8.9.24)$$

$$\Phi_{22} = -\dot{m}(u)/r^2. \quad (8.9.25)$$

Noticing (8.7.11a), especially $\Phi_{22} = R_{33}/2$ therein, we can see that the Ricci tensor of the Vaidya metric is

$$R_{ab} = R_{33}(\varepsilon^3)_a(\varepsilon^3)_b = R_{33}(-k_a)(-k_b) = 2\Phi_{22}k_ak_b = -2\dot{m}(u)r^{-2}(du)_a(du)_b, \quad (8.9.26)$$

which agrees with the R_{ab} [see (8.9.11)] derived using the coordinate basis method [Equation (3.4.21)].

Using the above result one can now also show that the matter field corresponding to the Vaidya metric is not an electromagnetic field. In the NP formalism, an electromagnetic field F_{ab} is represented by complex quantities Φ_0, Φ_1, Φ_2 , whose relations with $\Phi_{00}, \dots, \Phi_{22}$ representing the Ricci tensor are given in (8.8.8). Since Φ_{22} is the only nonvanishing one among $\Phi_{00}, \dots, \Phi_{22}$, (8.8.8) gives

$$\Phi_0 = \Phi_1 = 0, \quad \Phi_2 = Ae^{i\alpha}, \quad (8.9.27)$$

where $A \equiv \sqrt{-\dot{m}(u)/2r^{-1}}$, and α is a real function of the coordinates. Plugging (8.9.27) into the source-free Maxwell equations (8.8.3), one finds that (a), (c) are identities and (b), (d) leads to, respectively,

$$\frac{\partial\alpha}{\partial r} = 0, \quad -\frac{1}{\sin\theta}\frac{\partial\alpha}{\partial\varphi} - i\frac{\partial\alpha}{\partial\theta} = \cot\theta. \quad (8.9.28)$$

The first equation indicates that $\alpha = \alpha(u, \theta, \varphi)$, and the real and imaginary parts of the second equation gives $\partial\alpha/\partial\theta = 0$ [and hence $\alpha = \alpha(u, \varphi)$] and $\partial\alpha/\partial\varphi = -\cos\theta$. These two equations contradict each other. Thus, the matter field of the Vaidya metric is not an electromagnetic field, and therefore can only be a pure radiation field.

[The End of Optional Reading 8.9.1]

8.9.2 The Kinnersley Metric

The Vaidya metric is a generalization of the Schwarzschild metric, and a new metric defined by W. Kinnersley is a generalization of the Vaidya metric [Kinnersley (1969)]. Now we introduce this metric. Suppose $L(u)$ is an arbitrary smooth time-like curve (imagine it as the world line of a rocket) in 4-dimensional Minkowski spacetime $(\mathbb{R}^4, \eta_{ab})$, where u is the proper time. (Here we use u instead of τ , the purpose will be clear later). Following Kinnersley, we will use λ^a (instead of U^a in the convention of this text) to represent the 4-velocity of $L(u)$, i.e., $\lambda^a \equiv (\partial/\partial u)^a$. Suppose p is an arbitrary point in \mathbb{R}^4 , then L and the past light cone surface of p have exactly one intersection,¹² denoted by q (see Fig. 8.11). Let $\{X^\mu\}$ be an arbitrary inertial coordinate system, λ^μ be the components of λ^a in this system, and ψ^a, ξ^a be the position vectors of p, q in this system, i.e., $\psi^a \equiv \psi^\mu(\partial/\partial X^\mu)^a|_p$, $\xi^a \equiv \xi^\mu(\partial/\partial X^\mu)^a|_q$, where $\psi^\mu \equiv X^\mu(p)$, $\xi^\mu \equiv X^\mu(q)$. Originally, u and λ^a are

¹² There is an exception when $L(u)$ is asymptotically null (e.g., the hyperbola in Exercise 6.13). Kinnersley (1969) did not discuss this exception.

only a scalar field and a vector field defined on $L(u)$; however, their domains can be naturally extended to the whole \mathbb{R}^4 : $\forall p \in \mathbb{R}^4$, we have a unique $q \in L$, and thus we can define $u(p) := u(q)$, $\lambda^\mu(p) := \lambda^\mu(q)$. [Define $\lambda^a|_p$ by defining its coordinate components $\lambda^\mu(p)$, i.e., $\lambda^a|_p := \lambda^\mu(q)(\partial/\partial X^\mu)^a|_p$]. Thus, the parametric equations for each integral curve $C(u)$ of λ^a in the coordinate system $\{X^\mu\}$ are

$$X^\mu(u) = \xi^\mu(u) + \sigma^\mu \quad (\text{constants } \sigma^\mu \text{ satisfy } \eta_{\mu\nu}\sigma^\mu\sigma^\nu = 0). \quad (8.9.29)$$

[Because the tangent of the curve represented by the above parametric equations has components $dX^\mu(u)/du = d\xi^\mu(u)/du = \lambda^\mu$ in the system $\{X^\mu\}$. When $\sigma^\mu = 0$ the above equation will degenerate to $X^\mu(u) = \xi^\mu$, namely the parametric equations of $L(u)$]. This indicates that the λ^a of any point p satisfies

$$\lambda^a \partial_a u = (\partial/\partial u)^a \partial_a u = 1. \quad (8.9.30)$$

Define a vector $\sigma^a|_p := \psi^a - \xi^a$ at p . From Fig. 8.11 we can see that $\sigma^a|_p$ is null, and hence σ^a is a null vector field, which is the normal vector field of a family of null hypersurfaces (the family formed by the future light cone surfaces whose apices are points on L).

Since each point p has a timelike vector $\lambda^a|_p$ and a null vector $\sigma^a|_p$, one can apply the “3 + 1 decomposition” to $\sigma^a|_p$ and take $\lambda^a|_p$ as the time direction; that is, one can decompose σ^a into the sum of a component parallel to λ^a (denoted by $r\lambda^a$) and a component perpendicular to λ^a (denoted by $\hat{\sigma}^a$), i.e., (see Fig. 8.12)

$$\sigma^a = r\lambda^a + \hat{\sigma}^a. \quad (8.9.31)$$

Contracting both sides of this equation with $\lambda_a \equiv \eta_{ab}\lambda^b$, and noticing that $\lambda^a\lambda_a = -1$ and that $\hat{\sigma}^a$ is orthogonal to λ^a , we obtain

$$r = -\lambda_a\sigma^a. \quad (8.9.32)$$

Fig. 8.11 Each spacetime point p determines a point q on the timelike curve $L(u)$

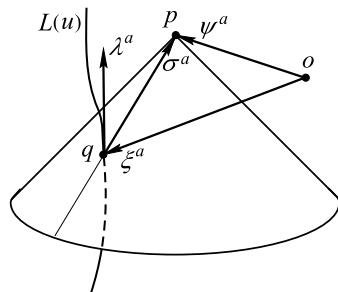
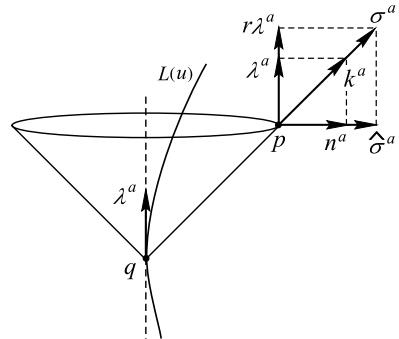


Fig. 8.12 The “3 + 1 decomposition” of a null vector σ^a (and k^a) at p



Also, let

$$k^a \equiv r^{-1}\sigma^a, \quad n^a \equiv r^{-1}\hat{\sigma}^a, \quad (8.9.33)$$

then we have

$$(a) \quad k^a = \lambda^a + n^a, \quad (b) \quad \lambda_a k^a = -1, \quad (c) \quad \eta_{ab} n^a n^b = 1. \quad (8.9.34)$$

k^a can be regarded as some kind of “normalization” of σ^a : the magnitudes of the time component λ^a and the spatial component n^a of k^a are both 1.

Since σ^a (and thus k^a) is a normal vector field on each future light cone surface with each point on L as the apex, $k_a \equiv \eta_{ab} k^b$ is the normal covector of each of these hypersurfaces. On the other hand, these hypersurfaces being constant- u surfaces indicates that $\partial_a u$ is their normal covector, and thus k_a and $\partial_a u$ at most differ by a multiplicative factor, i.e., $k_a = \alpha \partial_a u$. Combining this with (8.9.34)(b) and (8.9.30) yields $\alpha = -1$, and hence

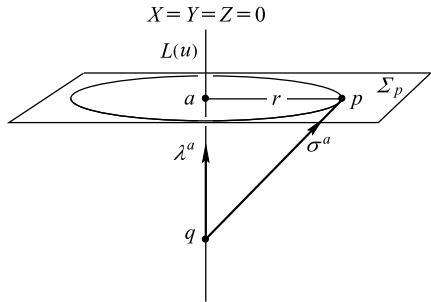
$$k_a = -(du)_a. \quad (8.9.35)$$

Based on the preceding discussion, Kinnersley defined a metric on \mathbb{R}^4 which later was dubbed with his last name. This metric can be expressed in terms of abstract indices as

$$g_{ab} := \eta_{ab} + 2m(u)r^{-1}k_a k_b = \eta_{ab} + 2m(u)r^{-1}(du)_a (du)_b, \quad (8.9.36)$$

where $m(u)$ is a function of u . Now that there are two metrics (η_{ab} and g_{ab}) on \mathbb{R}^4 [with $L(u)$ removed], we need to pay additional attention to raising and lowering indices (and other constructions involving a metric). For those quantities defined as vectors (each carries an upper index) in the first place, such as λ^a , σ^a and k^a , it is crystal clear. We stipulate that for all the tensors obtained by raising and lowering indices (e.g., λ_a , σ_a , k_a), the indices are raised and lowered by η_{ab} . For those tensors

Fig. 8.13 Choose the geodesic $L(u)$ as the world line of the spatial origin of the inertial frame $\{T, X, Y, Z\}$



whose indices are raised and lowered by g_{ab} we will write out g_{ab} explicitly, for instance $g_{ab}\lambda^b$ is not equal to $\lambda_a (= \eta_{ab}\lambda^b)$.¹³

First we discuss the simple case where $L(u)$ is a geodesic of η_{ab} (we will call it an η -geodesic for brevity). In this case, the Kinnersley metric (8.9.36) comes down to the Vaidya metric (when $\dot{m} \neq 0$) or the Schwarzschild metric (when $\dot{m} = 0$). In order to see this, one only needs to write out the line element of g_{ab} in an appropriate coordinate system $\{u, r, \theta, \varphi\}$ and compare it with (8.9.4). Take u and r which are already defined for each point as the first two coordinates of the system $\{u, r, \theta, \varphi\}$, and leave θ and φ to be defined below. Suppose $\{T, X, Y, Z\}$ is the inertial coordinate system of η_{ab} , whose origin of the spatial coordinates ($X = Y = Z = 0$) as a world line coincides with the geodesic $L(u)$, then the components of λ^a in this coordinates are $\lambda^\mu = (1, 0, 0, 0)$ (see Fig. 8.13), and hence the r in (8.9.32) satisfies

$$r = -\eta_{\mu\nu}\sigma^\mu\lambda^\nu = -\eta_{00}\sigma^0\lambda^0 = \sigma^0.$$

On the other hand, the 3-dimensional space Σ_p in Fig. 8.13 can be viewed as the whole space at the time of p . From the figure we can see that σ^0 = the length of the line segment qa = the length of the line segment ap , and thus $r = \sigma^0$ indicates that the value of r at p is the spatial distance between p and the geodesic $L(u)$. Set up a spherical coordinate system $\{r, \theta, \varphi\}$ on Σ_p with a as the origin and r as the radial coordinate, in which θ and φ are defined as follows:

$$X = r \sin \theta \cos \varphi, \quad Y = r \sin \theta \sin \varphi, \quad Z = r \cos \theta.$$

Combining this $\{r, \theta, \varphi\}$ with u yields the 4-dimensional coordinate system we want. The u and r of this system and the T of $\{T, X, Y, Z\}$ has the following relation: $T = u + r$. Hence, the line element of η_{ab} in this system is $-du^2 - 2dudr + r^2(d\theta^2 + \sin^2 \theta d\varphi^2)$, and therefore the line element of the Kinnersley metric g_{ab} is

$$ds^2 = -[1 - 2m(u)r^{-1}]du^2 - 2dudr + r^2(d\theta^2 + \sin^2 \theta d\varphi^2), \quad (8.9.37)$$

¹³ However, $g_{ab}k^b$ is equal to $k_a (= \eta_{ab}k^b)$. This is because it follows from (8.9.36) that $g_{ab}k^b = \eta_{ab}k^b + 2mr^{-1}(du)_a(du)_bk^b$, and $k^b(du)_b = k^b\partial_b u = 0$ (According to the definition of u , it is a constant on an integral curve of k^a).

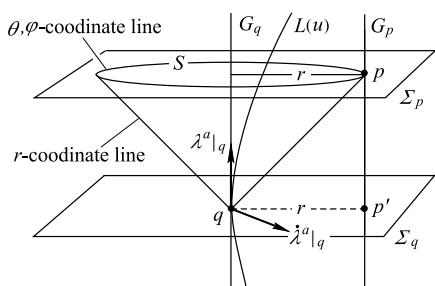
which has the same form as (8.9.4). Thus, the Kinnersley metric (8.9.36) comes down to the Vaidya metric (when $\dot{m} \neq 0$) or the Schwarzschild metric (when $\dot{m} = 0$) in the case where $L(u)$ is an η -geodesic. The actual generalization by Kinnersley is the case where $L(u)$ is not an η -geodesic, which will be discussed in detail below.

8.9.3 The Kinnersley Metric (Detailed Discussions)

When $L(u)$ is not an η -geodesic, the 4-acceleration $\lambda^b \partial_b \lambda^a$ of $L(u)$ is nonvanishing (∂_b is the derivative operator associated with η_{ab}), and is orthogonal to the 4-velocity λ^a . Following Kinnersley, we use $\dot{\lambda}^a$ to represent the 4-acceleration $\lambda^b \partial_b \lambda^a$, then $\eta_{ab} \lambda^a \dot{\lambda}^b = 0$. We still want to choose an appropriate coordinate system $\{u, r, \theta, \varphi\}$ to represent the Kinnersley metric. The definitions of u and r are the same as those in the case where $L(u)$ is a geodesic, just the geometric interpretation of r is slightly different now. Suppose p is an arbitrary spacetime point, then it determines a unique point q on $L(u)$. Let G_q be the η -geodesic passing through q , and denote the inertial reference frame determined by G_q by \mathcal{R}_q . Let Σ_p and Σ_q represent two surfaces of simultaneity in \mathcal{R}_p , which respectively include p and q . Following the discussion in Sect. 8.9.2 we can see that the r determined by (8.9.32) represents the spatial distance between p and G_q (see Fig. 8.14). Let G_p represent the inertial observer that passes through p in \mathcal{R}_q , then r can also be viewed as the spatial distance between G_p and $L(u)$ at the moment Σ_p (i.e., the distance between p' and q in the figure). In electrodynamics, Σ_q is referred to as the “retarded time” corresponding to Σ_p (note that Σ_p is actually later than Σ_q , but Σ_q is conventionally called the retarded time), and hence what r stands for is the *retarded* distance between G_p and $L(u)$.

Now we will introduce the definition of the coordinate θ and φ in the system $\{u, r, \theta, \varphi\}$. Any point q on L determines an instantaneous rest inertial observer G_q and an instantaneous rest inertial reference frame \mathcal{R}_q . Define an instantaneous rest inertial coordinate system $\{X^\mu\} \equiv \{T, X, Y, Z\}$ in \mathcal{R}_q based on the following requirements: ① take G_q as the world line of the origin of the spatial coordinates ($X = Y = Z = 0$), ② take q as the point at $T = 0$ on this world line, ③ the Z -axis and $\dot{\lambda}^a|_q$ are in the same direction (the direction of the X -axis still has arbitrariness).

Fig. 8.14 The value of r at p gives the retarded distance between the observer G_p and the rocket $L(u)$



Take the Z -axis of this system as the polar axis, define the coordinates θ and φ on the future light cone surface of q (including q) as follows:

$$X = r \sin \theta \cos \varphi, \quad Y = r \sin \theta \sin \varphi, \quad Z = r \cos \theta. \quad (8.9.38)$$

Since the direction of the 4-acceleration $\dot{\lambda}^a$ changes continuously as q moves along L , when defining θ and φ we need to keep rotating the direction pointing at the north pole in order to guarantee that it keeps align with $\dot{\lambda}^a$.

By a calculation based on the preceding discussion (see Optional Reading 8.9.2 for details) one can find all the nonvanishing components of the Kinnersley metric g_{ab} in the system $\{u, r, \theta, \varphi\}$:

$$\begin{aligned} g_{uu} &= -1 - 2a(u)r \cos \theta + r^2(f^2 + g^2 \sin^2 \theta) + 2m(u)r^{-1}, & g_{ur} = g_{ru} &= -1, \\ g_{u\theta} = g_{\theta u} &= -r^2 f, & g_{u\varphi} = g_{\varphi u} &= -r^2 g \sin^2 \theta, & g_{\theta\theta} &= r^2, & g_{\varphi\varphi} &= r^2 \sin^2 \theta, \end{aligned} \quad (8.9.39a)$$

with

$$f \equiv a(u) \sin \theta + b(u) \sin \varphi - c(u) \cos \varphi, \quad g \equiv [b(u) \cos \varphi + c(u) \sin \varphi] \cot \theta, \quad (8.9.39b)$$

where

$$a(u) \equiv |\dot{\lambda}^a(u)| \equiv [\eta_{ab} \dot{\lambda}^a(u) \dot{\lambda}^b(u)]^{1/2} \quad (8.9.39c)$$

is the magnitude of the 4-acceleration of $L(u)$,¹⁴ and b and c describe the time rate of change (u as the time) of the direction of $\dot{\lambda}^a$, see Optional Reading 8.9.2 for details. If a segment of $L(u)$ is a timelike hyperbola (see Exercise 6.13), then $a = \text{constant}$ and $b = c = 0$ in this segment.

One can further calculate the Ricci tensor R_{ab} and scalar curvature R of the Kinnersley metric:

$$R_{ab} = -2r^{-2}(\dot{m} + 3ma \cos \theta)k_a k_b, \quad R = 0, \quad (8.9.40)$$

and hence the corresponding T_{ab} is

$$T_{ab} = -\frac{1}{4\pi r^2}(\dot{m} + 3ma \cos \theta)k_a k_b. \quad (8.9.41)$$

Similar to (8.9.12'), the matter field corresponding to the above expression is also a pure radiation field rather than an electromagnetic field. Although this matter field is formed by massless particles which are not photons, we will refer to them as “photons” for the sake of convenience.

¹⁴ Note that the $a(u)$ defined in this text has a sign difference compared with Kinnersley (1969) and Bonnor (1994).

As we have mentioned, the Kinnersley metric comes down to the Vaidya metric when $L(u)$ is an η -geodesic and $\dot{m} \neq 0$. By means of $L(u)$ we can provide a more intuitive interpretation for the physical meaning of the Vaidya metric. In this interpretation one should note that there are two metric fields on \mathbb{R}^4 , namely η_{ab} and the Vaidya metric $g_{ab}^{(Vai)}$; the geodesic, 4-acceleration, etc. we mentioned above are all measured by η_{ab} and its associated derivative operator ∂_a .

One may imagine this: a star is undergoing geodesic motion (inertial motion) in Minkowski space with $L(u)$ as its world line. Since it keeps emitting particles, its mass (energy) keeps decreasing ($\dot{m} < 0$). The 4-momentum of the star together with the energy-momentum T_{ab} of the surrounding radiation field produce a gravitational field which makes the spacetime curved, and the spacetime is described by $g_{ab}^{(Vai)}$. [However, one cannot ask a question like “is the world line a geodesic measured by $g_{ab}^{(Vai)}$ ”, since $g_{ab}^{(Vai)}$ is not defined on the curve ($r = 0$)]. Since the geodesic $L(u)$ “holds the scales even”, i.e., it is isotropic, $g_{ab}^{(Vai)}$ has spherical symmetry, but $\dot{m} \neq 0$ makes it lose the stationarity. This intuitive physical interpretation can also be carried over to the Kinnersley metric $g^{(Kin)}$. Now $L(u)$ is not an η -geodesic, and its radiation is not isotropic anymore; hence, it is not appropriate to regard $L(u)$ as the world line of a star. Thus, we now change the star to a rocket, which keeps emitting “photons” outwards in an anisotropic manner (to some extent similar to a real rocket emitting jets), and hence is called a “photon rocket” in the literature. The recoil experienced by this rocket due to the fact that it emits photons makes its energy and 3-momentum keep changing; the former is manifested by $\dot{m} < 0$, and the latter renders the time rate of change of the 3-momentum nonvanishing. Formulating in the 4-dimensional language, using P^a to represent the 4-momentum of the rocket, we have $P^a = m\lambda^a$, and hence its time rate of change is $\dot{P}^a = \dot{m}\lambda^a + m\dot{\lambda}^a$, where $\dot{P}^a \equiv \lambda^b \partial_b P^a$. The first and second terms represent the time rates of change of the energy and 3-momentum, respectively. In the instantaneous rest inertial frame $\{X^\mu\}$ at q , the component expression for this equation reads

$$\dot{P}^\mu = \dot{m}\lambda^\mu + m\dot{\lambda}^\mu. \quad (8.9.42)$$

Since at q we have

$$\lambda^\mu = (1, 0, 0, 0), \quad \dot{\lambda}^\mu = (0, 0, 0, a) \quad (\dot{\lambda}^a \text{ is in the direction of the } Z\text{-axis}), \quad (8.9.43)$$

the rocket has the following increasing rates:

$$\text{increasing rate of the energy} = \dot{P}^0 = \dot{m}\lambda^0 = \dot{m}, \quad (8.9.44a)$$

$$\text{increasing rate of the } i\text{-component of the momentum} = \dot{P}^i = m\dot{\lambda}^i = (0, 0, ma). \quad (8.9.44b)$$

Now we will show that the energy and momentum increasing rates of the rocket caused by this recoil are exactly the energy and momentum carried by the “photons” emitted by the rocket to infinity per unit time times -1 . To do so, we should calculate

the energy and momentum flowing out of the sphere S in Fig. 8.14. Suppose $\{X^\mu\}$ is an instantaneous rest inertial frame at q , then $\{(e_\mu)^a\} \equiv \{(\partial/\partial X^\mu)^a\}$ is an orthonormal tetrad field on \mathbb{R}^4 . Sect. 6.4 points out that $T^{0j} (= -T_{0j})$ is the j -component of the energy flux density, and hence $T^{0j}(e_j)^a$ is the energy flux density vector. Therefore,

$$\text{energy flowing outside } S \text{ per unit time} = \int_S T^{0j}(e_j)^a n_a dS = \int_S T^{0j} n_j dS, \quad (8.9.45)$$

where $n_a \equiv \eta_{ab} n^b$, while n^b is the outgoing unit normal vector of the sphere S , namely the n^a in (8.9.33). Moreover, Sect. 6.4 also points out that $T^{ij}(e_i)^a(e_j)^b$ is the 3-momentum flux density tensor, whose contraction with any spatial unit vector gives the 3-momentum flux density vector. Therefore,

$$\text{3-momentum flowing out of } S \text{ per unit time} = \int_S T^{ij}(e_i)^a(e_j)^b n_b dS = \int_S T^{ij}(e_i)^a n_j dS, \quad (8.9.46)$$

$$i\text{-component of the 3-momentum flowing out of } S \text{ per unit time} = \int_S T^{ij} n_j dS. \quad (8.9.47)$$

Summarizing (8.9.45) and (8.9.47) we can write

$$\mu\text{-component of the 4-momentum flowing out of } S \text{ per unit time} = \int_S T^{\mu\nu} n_\nu dS. \quad (8.9.48)$$

It follows from the definition of the instantaneous rest inertial frame $\{X^\mu\}$ at q and the coordinates θ and φ that at any point on S we have (see Figs. 8.14 and 8.12)

$$\lambda^\mu = (1, 0, 0, 0) \quad \text{and} \quad n^\mu = (0, \sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta). \quad (8.9.49)$$

Plugging in $k^a = \lambda^a + n^a$ yields

$$k^\mu = (1, \sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta), \quad (8.9.50)$$

and thus $k^\mu n_\mu = 1$. Combining this with (8.9.41) yields

$$T^{\mu\nu} n_\nu = -\frac{1}{4\pi r^2} (\dot{m} + 3ma \cos \theta) k^\mu k^\nu n_\nu = -\frac{1}{4\pi r^2} (\dot{m} + 3ma \cos \theta) k^\mu, \quad (8.9.51)$$

and hence

$$\begin{aligned} \text{energy flowing out of } S \text{ per unit time} &= -\frac{1}{4\pi r^2} \int_S (\dot{m} + 3ma \cos \theta) k^0 dS \\ &= -\frac{1}{4\pi r^2} \int_0^{2\pi} d\varphi \int_0^\pi (\dot{m} + 3ma \cos \theta) r^2 \sin \theta d\theta = -\dot{m}, \end{aligned} \quad (8.9.52a)$$

the 3rd component of the 3-momentum flowing out of S per unit time

$$\begin{aligned} &= -\frac{1}{4\pi r^2} \int_S (\dot{m} + 3ma \cos \theta) k^3 dS \\ &= -\frac{1}{4\pi r^2} \int_0^{2\pi} d\varphi \int_0^\pi (\dot{m} + 3ma \cos \theta) \cos \theta r^2 \sin \theta d\theta = -ma. \end{aligned} \quad (8.9.52b)$$

Similarly we obtain that

$$\text{the 1st and 2nd components of the 3-momentum flowing out of } S \text{ per unit time} = 0. \quad (8.9.52c)$$

Equations (8.9.52) also hold when $r \rightarrow \infty$. Comparing them with (8.9.44) proves the conclusion we claimed above, i.e., the increasing rates of the rocket's energy and momentum are exactly the energy and momentum carried by the "photons" it emits to infinity per unit time times -1 .

Based on the physical interpretation above, we may refer to the Kinnersley solution as the "solution of an arbitrary accelerating point mass", or say that the Kinnersley metric represents the "gravitational field of an arbitrary accelerating point mass". However, one should note that: ① "accelerating point mass" means that the 4-acceleration $\dot{\lambda}^a \equiv \lambda^b \partial_b \lambda^a$ of the rocket is nonvanishing ($a \neq 0$), and this 4-acceleration is measured by η_{ab} . Why is it not measured by g_{ab} ? The answer is: the world line of the rocket has $r = 0$, while g_{ab} is not well-defined (is singular) on this curve, and so it cannot be used to measure any quantity on the rocket's world line. ② This "gravitational field of an accelerating point mass" is generated by the point mass (rocket) together with the "photons" it emits, and the T_{ab} corresponding to g_{ab} is the energy-momentum tensor of the pure radiation field outside the rocket.

The preceding discussion about the Kinnersley metric also has a few subtleties, as we will list below:

(1) When computing the energy and momentum flowing out of the sphere S , we have used η_{ab} for everything that involves a metric without mentioning; however, the metric of Kinnersley spacetime is supposed to be the Kinnersley metric, and so the legitimacy of the above calculation should be called into question. Regarding this, Bonnor (1994) provides an answer as follows (gist, not exact words): the difference between g_{ab} and η_{ab} is only in the term with mr^{-1} . Adding this term will affect the normalization of the n_v in (8.9.51), but its contribution to the integral will approach zero when S approaches infinity. Thus, it turns out that ignoring the term with mr^{-1} will not affect the upshot.

(2) For any matter field known by physicists, the energy density measured by any observer at any time is non-negative (called the weak energy condition, see Appendix D in Volume II for details). Suppose (p, Z^a) is an arbitrary instantaneous observer, then it follows from (8.9.45) that

$$T_{00} = T_{ab} Z^a Z^b = -\frac{1}{4\pi r^2} (k_a Z^a)^2 (\dot{m} + 3ma \cos \theta).$$

When $a = 0$ (Vaidya), we only have to let $\dot{m} < 0$ to guarantee $T_{00} > 0$. However, the case where $a \neq 0$ is not that simple since $\cos \theta$ can be both positive and negative. Nevertheless, as long as we assume $m > 0$, it is not difficult to see that $T_{00} > 0$ is equivalent to $-\dot{m}/3m \geq a \cos \theta$. Therefore, in order to make T_{00} non-negative for any value of θ , besides $\dot{m} < 0$, we should also require that $a \leq -\dot{m}/3m$. One can consider this as some sort of constraint coming from the energy condition on the relation between the two parameters m and a of the Kinnersley metric.

(3) Bonnor (1994) points out that, since the rocket undergoes an accelerating motion, it should emit gravitational waves which carry energy and momentum out to infinity. However, we have proved that under the premise without gravitational waves, the energy and momentum carried only by the “photons” to infinity have already satisfied the balance requirement, i.e., they are exactly the energy and momentum increasing rates times -1 . This implies that the energy and momentum carried by gravitational waves to infinity vanishes. Hence, there is a paradox: does the Kinnersley spacetime have any gravitational radiation at all? Regarding this problem, Damour (1995) and Dain et al. (1996) studied the gravitational radiation of the Kinnersley metric using very different approaches, and the basic conclusion is: both the point-like accelerating rocket and the “photons” it is surrounded by emit gravitational radiation; the energy and momentum carried by them cancel each other, and so overall there are no gravitational waves in Kinnersley spacetime (the energy and momentum carried by the gravitational waves to infinity vanish).

[Optional Reading 8.9.2]

Now we provide the detailed derivation of (8.9.39). It follows from (8.9.36) that among all the components of g_{ab} and η_{ab} in the system $\{u, r, \theta, \varphi\}$ the only different one is the uu -component. Specifically speaking, if we use $g_{uu}, g_{ur}, \dots, g_{\varphi\varphi}$ and ${}^0g_{uu}, {}^0g_{ur}, \dots, {}^0g_{\varphi\varphi}$ to represent the components of g_{ab} and η_{ab} in the system $\{u, r, \theta, \varphi\}$, then

$$\begin{aligned} g_{uu} &= {}^0g_{uu} + 2mr^{-1}, & g_{ur} &= {}^0g_{ur}, & g_{u\theta} &= {}^0g_{u\theta}, & g_{u\varphi} &= {}^0g_{u\varphi}, \\ g_{rr} &= {}^0g_{rr}, & g_{r\theta} &= {}^0g_{r\theta}, & g_{r\varphi} &= {}^0g_{r\varphi}, & g_{\theta\theta} &= {}^0g_{\theta\theta}, & g_{\theta\varphi} &= {}^0g_{\theta\varphi}, & g_{\varphi\varphi} &= {}^0g_{\varphi\varphi}. \end{aligned} \quad (8.9.53)$$

Therefore, we only have to compute ${}^0g_{uu}, {}^0g_{ur}, \dots, {}^0g_{\varphi\varphi}$.

Let us compute ${}^0g_{uu}|_p, {}^0g_{ur}|_p, \dots, {}^0g_{\varphi\varphi}|_p$ for an arbitrary point p in \mathbb{R}^4 . A point p determines a point q on L , and we have defined an instantaneous rest inertial coordinate system $\{X^\mu\} \equiv \{T, X, Y, Z\}$ by means of q . Denoting the ψ^μ in $\sigma^\mu = \psi^\mu - \xi^\mu$ by X^μ yields $X^\mu = \sigma^\mu + \xi^\mu$. Then using (8.9.33) we obtain the coordinate transformation between the two systems:

$$X^\mu = \sigma^\mu + \xi^\mu = rk^\mu(u, \theta, \varphi) + \dot{\xi}^\mu(u), \quad (8.9.54)$$

where k^μ represents the components of k^a in the system $\{X^\mu\}$. [Any quantity with indices μ, ν, \dots or $0, 1, \dots$ represents the components of a certain tensor in the system $\{X^\mu\}$ (not $\{u, r, \theta, \varphi\}$)]. Since $\{X^\mu\}$ is an inertial coordinate system, the components of η_{ab} in $\{X^\mu\}$ are certainly $\eta_{\mu\nu}$, and using the coordinate transformation (8.9.54) one can write down the expressions for the components of η_{ab} in the system $\{u, r, \theta, \varphi\}$. First,

$$\begin{aligned} {}^0g_{uu} &= \eta_{\mu\nu} \frac{\partial X^\mu}{\partial u} \frac{\partial X^\nu}{\partial u} = \eta_{\mu\nu} (r\dot{k}^\mu + \dot{\xi}^\mu)(r\dot{k}^\nu + \dot{\xi}^\nu) \\ &= r^2 \eta_{\mu\nu} \dot{k}^\mu \dot{k}^\nu + 2r \eta_{\mu\nu} \dot{k}^\mu \dot{\xi}^\nu + \eta_{\mu\nu} \dot{\xi}^\mu \dot{\xi}^\nu, \end{aligned}$$

where the dotted quantities stand for the (partial) derivatives with respect to u , e.g., $\dot{\xi}^0 \equiv d\xi^0/du$, $\dot{k}^1 \equiv dk^1/\partial u$. Since the parametric equations of the curve $L(u)$ are $X^\mu(u) = \xi^\mu(u)$, $\dot{\xi}^\mu \equiv d\xi^\mu/du$ is equal to the components λ^μ of the tangent vector λ^a of $L(u)$ in the system $\{X^\mu\}$. From $\eta_{\mu\nu}\lambda^\mu\lambda^\nu = -1$ we obtain

$${}^0g_{uu} = -1 + 2r\eta_{\mu\nu}\dot{k}^\mu\lambda^\nu + r^2\eta_{\mu\nu}\dot{k}^\mu\dot{k}^\nu. \quad (8.9.55a)$$

Second,

$${}^0g_{ur} = \eta_{\mu\nu}\frac{\partial X^\mu}{\partial u}\frac{\partial X^\nu}{\partial r} = \eta_{\mu\nu}(r\dot{k}^\mu + \dot{\xi}^\mu)k^\nu = r\eta_{\mu\nu}\dot{k}^\mu k^\nu + \eta_{\mu\nu}\lambda^\mu k^\nu = -1, \quad (8.9.55b)$$

where $\eta_{\mu\nu}\dot{k}^\mu k^\nu = 0$ can be derived from $\eta_{\mu\nu}k^\mu k^\nu = 0$, while $\eta_{\mu\nu}\lambda^\mu k^\nu = -1$ comes from $\lambda_a k^a = -1$. In a similar manner one can find the expressions for the other components of η_{ab} in $\{u, r, \theta, \varphi\}$:

$${}^0g_{u\theta} = r^2\eta_{\mu\nu}\dot{k}^\mu k^\nu,_\theta + r\eta_{\mu\nu}\lambda^\mu k^\nu,_\theta, \quad (8.9.55c)$$

$${}^0g_{u\varphi} = r^2\eta_{\mu\nu}\dot{k}^\mu k^\nu,_\varphi + r\eta_{\mu\nu}\lambda^\mu k^\nu,_\varphi, \quad (8.9.55d)$$

$${}^0g_{rr} = \eta_{\mu\nu}k^\mu k^\nu = 0, \quad (8.9.55e)$$

$${}^0g_{r\theta} = r\eta_{\mu\nu}k^\mu k^\nu,_\theta = 0, \quad (8.9.55f)$$

$${}^0g_{r\varphi} = r\eta_{\mu\nu}k^\mu k^\nu,_\varphi = 0, \quad (8.9.55g)$$

$${}^0g_{\theta\theta} = r^2\eta_{\mu\nu}k^\mu,_\theta k^\nu,_\theta, \quad (8.9.55h)$$

$${}^0g_{\theta\varphi} = r^2\eta_{\mu\nu}k^\mu,_\theta k^\nu,_\varphi, \quad (8.9.55i)$$

$${}^0g_{\varphi\varphi} = r^2\eta_{\mu\nu}k^\mu,_\varphi k^\nu,_\varphi, \quad (8.9.55j)$$

where the second equalities in (f) and (g) come from $\eta_{\mu\nu}k^\mu k^\nu = 0$. In order to find the final form of the above expressions, one must compute the partial derivatives of k^μ with respect to u , θ and φ , i.e., \dot{k}^μ , $k^\mu,_\theta$ and $k^\mu,_\varphi$. To find $k^\mu,_\theta$ and $k^\mu,_\varphi$, one only needs to care about the k^μ on the future light cone surface with the fixed q being the apex. In this case (8.9.50) holds, and we can again list the following expression (with a new equation number):

$$k^\mu = (1, \sin\theta \cos\varphi, \sin\theta \sin\varphi, \cos\theta), \quad (8.9.56a)$$

and hence

$$k^\mu,_\theta = (0, \cos\theta \cos\varphi, \cos\theta \sin\varphi, -\sin\theta), \quad (8.9.56b)$$

$$k^\mu,_\varphi = (0, -\sin\theta \sin\varphi, \sin\theta \cos\varphi, 0). \quad (8.9.56c)$$

Thus,

$$\eta_{\mu\nu}k^\mu,_\theta k^\nu,_\theta = 1, \quad \eta_{\mu\nu}k^\mu,_\theta k^\nu,_\varphi = 0, \quad \eta_{\mu\nu}k^\mu,_\varphi k^\nu,_\varphi = \sin^2\theta. \quad (8.9.56d)$$

Moreover, $\lambda^\mu = (1, 0, 0, 0)$ also leads to

$$\eta_{\mu\nu}\lambda^\mu k^\nu,_\theta = \eta_{\mu\nu}\lambda^\mu k^\nu,_\varphi = 0. \quad (8.9.56e)$$

Now we have the most complicated step remaining, namely computing \dot{k}^μ .

Suppose p and \tilde{p} are two neighboring spacetime points, whose values of r, θ, φ are the same, while the values of u are respectively u and $u + du$. Let q and \tilde{q} represent the points corresponding to p and \tilde{p} on $L(u)$, then $k^a|_p$ points from q to p and $k^a|_{\tilde{p}}$ points from \tilde{q} to \tilde{p} . Denote $k^a \equiv k^a|_p$, $\chi^a \equiv k^a|_{\tilde{p}}$, then

$$\dot{k}^\mu|_p = \lim_{du \rightarrow 0} \frac{\chi^\mu - k^\mu}{du}, \quad (8.9.57)$$

where k^μ and χ^μ are respectively the components of k^a and χ^a in the instantaneous rest inertial coordinate system $\{X^\mu\} \equiv \{T, X, Y, Z\}$ at q . Now that k^μ has already been expressed as (8.9.56a), the main thing is how to derive χ^μ . Let $\{\tilde{X}^\mu\} \equiv \{\tilde{T}, \tilde{X}, \tilde{Y}, \tilde{Z}\}$ represent the instantaneous rest inertial coordinate system at \tilde{q} [according to the definition in the paragraph containing (8.9.38), one only needs to change q to \tilde{q}], then the components of χ^a in the system $\{\tilde{X}^\mu\}$ are

$$\tilde{\chi}^\mu = (1, \sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta). \quad (8.9.58)$$

To derive $\tilde{\chi}^\mu$ from χ^μ , we should first clarify the relation between the systems $\{\tilde{X}^\mu\}$ and $\{X^\mu\}$. According to our requirement, the Z -axis in $\{X^\mu\}$ should be aligned with the direction of $\dot{\lambda}^a|_q$, and the \tilde{Z} -axis in $\{\tilde{X}^\mu\}$ should be aligned with the direction of $\dot{\lambda}^a|_{\tilde{q}}$. Note that $\{\tilde{X}^\mu\}$ and $\{X^\mu\}$ are inertial coordinate systems in two different inertial reference frames $\mathcal{R}_{\tilde{q}}$ and \mathcal{R}_q , since the T -coordinate line G_q and the \tilde{T} -coordinate line $G_{\tilde{q}}$ [the η -geodesic tangent to $L(u)$] are not parallel in general. However, since both of them are inertial coordinate system, one can always transfer one to the other via an appropriate translation and Lorentz transformation. This transformation can be realized by the following three steps: ① Transfer the origin of $\{X^\mu\}$ (namely the point with $T = X = Y = Z = 0$) from q to \tilde{q} and obtain a coordinate system $\{X'^\mu\}$. ② Use a boost in the $T'Z'$ -plane to transfer $\{X'^\mu\}$ to another system $\{\hat{X}^\mu\}$ (where the \hat{T} -axis is parallel to the \tilde{T} -axis). This is an inertial coordinate system in the inertial reference frame $\mathcal{R}_{\tilde{q}}$ just like $\{\tilde{X}^\mu\}$, only the \hat{Z} -axis is in general not parallel to $\dot{\lambda}^a|_{\tilde{q}}$, which is the key difference between $\{\hat{X}^\mu\}$ and $\{\tilde{X}^\mu\}$. ③ Apply a spatial rotation R to $\{\hat{X}^\mu\}$ and turn it into $\{\tilde{X}^\mu\}$, in which the \tilde{Z} -axis is aligned with $\dot{\lambda}^a|_{\tilde{q}}$. This R can be considered as two rotations R_1 and R_2 acting successively (a composite map). R_1 is a rotation around the \hat{X} -axis that turns the \hat{Z} -axis to a new position (denoted by \hat{Z} , see Fig. 8.15), which is the intersection of the $\hat{Y}\hat{Z}$ -plane and the cone with the \hat{Y} -axis as the axis and \tilde{Z} as a generatrix; R_2 is a rotation around the \hat{Y} -axis that turns the \hat{Z} -axis to the \tilde{Z} -axis. Suppose the angles for R_1 and R_2 are bdu and cdu .¹⁵ These three steps can be expressed as

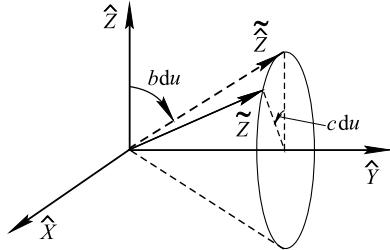
$$\{T, X, Y, Z\} \xrightarrow{\text{translation}} \{T', X', Y', Z'\} \xrightarrow{\text{boost}} \{\hat{T}, \hat{X}, \hat{Y}, \hat{Z}\} \xrightarrow{\text{spatial rotation } R} \{\tilde{T}, \tilde{X}, \tilde{Y}, \tilde{Z}\}. \quad (8.9.59)$$

The necessity of the spatial rotation R comes from our requirement that the polar axis (Z -axis) of θ and φ is always aligned with the direction of $\dot{\lambda}^a$. If $\dot{\lambda}^a|_{\tilde{q}}$ and $\dot{\lambda}^a|_q$ are parallel (which means the direction of $\dot{\lambda}^a$ does not change in time du), then since the boost in the $T'Z'$ -plane preserves the directions of X' , Y' and Z' , we can assure that the \hat{Z} -axis is aligned with the direction of $\dot{\lambda}^a|_{\tilde{q}}$ without another spatial rotation, and thus $b = c = 0$. Conversely, as long as $\dot{\lambda}^a|_{\tilde{q}}$ and $\dot{\lambda}^a|_q$ are not parallel, then the \hat{Z} -axis will not be aligned with the direction of $\dot{\lambda}^a|_{\tilde{q}}$, and hence we must rotate it by bdu and cdu so that the \tilde{Z} -axis is in the direction of $\dot{\lambda}^a|_{\tilde{q}}$. Thus, b and c indeed reflect the rate of change of the direction of the 4-acceleration $\dot{\lambda}^a$.

Having the ideas above, one can compute χ^μ from the expression (8.9.58) of $\tilde{\chi}^\mu$, and then derive \dot{k}^μ by plugging the result into (8.9.57). Since the spatial coordinate systems $\{\hat{X}, \hat{Y}, \hat{Z}\}$ and $\{\tilde{X}, \tilde{Y}, \tilde{Z}\}$ are related by a spatial rotation R , the components of χ^a in these two systems, $\hat{\chi}^i$ and $\tilde{\chi}^i$, can be expressed in terms of column matrices satisfying the following equation:

¹⁵ After these two rotations, the X -axis may still not be coincide with the \tilde{X} -axis, but this is not a problem since the choice of the X -axis if the instantaneous rest inertial frame at each point is flexible. One should "foresee" this and choose the \tilde{X} -axis based on the result of rotating the X -axis.

Fig. 8.15 After rotating around the \hat{X} -axis by bdu , and then rotating around the \hat{Y} -axis by cdu , the \hat{Z} -axis will turn into the \tilde{Z} -axis



$$\begin{bmatrix} \hat{\chi}^1 \\ \hat{\chi}^2 \\ \hat{\chi}^3 \end{bmatrix} = R \begin{bmatrix} \tilde{\chi}^1 \\ \tilde{\chi}^2 \\ \tilde{\chi}^3 \end{bmatrix}, \quad (8.9.60)$$

where $R = R_2 R_1$ is the 3×3 matrix described by the rotating angles bdu and cdu . From Fig. 8.15 and Appendix G in Volume II we have

$$\begin{aligned} R = R_2 R_1 &= \begin{bmatrix} \cos(cdu) & 0 & \sin(cdu) \\ 0 & 1 & 0 \\ -\sin(cdu) & 0 & \cos(cdu) \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(bdu) & -\sin(bdu) \\ 0 & \sin(bdu) & \cos(bdu) \end{bmatrix} \\ &= \begin{bmatrix} \cos(cdu) & \sin(bdu) \sin(cdu) & \cos(bdu) \sin(cdu) \\ 0 & \cos(bdu) & -\sin(bdu) \\ -\sin(cdu) & \sin(bdu) \cos(cdu) & \cos(bdu) \cos(cdu) \end{bmatrix}. \end{aligned}$$

Since du will approach zero at last, one can take $\cos(bdu) \cong \cos(cdu) \cong 1$, $\sin(bdu) \cong bdu$, $\sin(cdu) \cong cdu$. Ignoring the 2nd-order small terms containing $(du)^2$, we obtain

$$R = \begin{bmatrix} 1 & 0 & cdu \\ 0 & 1 & -bdu \\ -cdu & bdu & 1 \end{bmatrix}. \quad (8.9.61)$$

Plugging the above equation and $\tilde{\chi}^i$ given by (8.9.58) into (8.9.60) yields

$$\begin{aligned} \begin{bmatrix} \hat{\chi}^1 \\ \hat{\chi}^2 \\ \hat{\chi}^3 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & cdu \\ 0 & 1 & -bdu \\ -cdu & bdu & 1 \end{bmatrix} \begin{bmatrix} \sin \theta \cos \varphi \\ \sin \theta \sin \varphi \\ \cos \theta \end{bmatrix} \\ &= \begin{bmatrix} \sin \theta \cos \varphi + cdu \cos \theta \\ \sin \theta \sin \varphi - bdu \cos \theta \\ -cdu \sin \theta \cos \varphi + bdu \sin \theta \sin \varphi + \cos \theta \end{bmatrix}. \quad (8.9.62a) \end{aligned}$$

Since the spatial rotation does not affect the 0-component of a 4-vector, we have

$$\hat{\chi}^0 = \tilde{\chi}^0 = 1 \quad [(8.9.58) \text{ is used in the second equality}], \quad (8.9.62b)$$

and combining this with (8.9.62a) we obtain all $\hat{\chi}^\mu$. However, we want χ^μ in (8.9.57). It follows from (8.9.59) that $\{\hat{X}^\mu\}$ and $\{X^\mu\}$ are related by a translation and a boost; since the translation does not affect the components of a 4-vector, we only have to consider the effect of the boost. Since $\{\hat{X}^\mu\}$ is moving along the Z -axis relative to $\{X^\mu\}$ with a speed $v \equiv adu$ and $du \rightarrow 0$ assures that $\gamma \equiv (1 - v^2)^{-1/2} \rightarrow 1$, from the Lorentz transformation with $\gamma \cong 1$ we get

$$\chi^0 = \hat{\chi}^0 + (adu)\hat{\chi}^3, \quad \chi^1 = \hat{\chi}^1, \quad \chi^2 = \hat{\chi}^2, \quad \chi^3 = \hat{\chi}^3 + (adu)\hat{\chi}^0. \quad (8.9.63)$$

Plugging the $\hat{\chi}^\mu$ of (8.9.62) into the equation above gives

$$\begin{aligned}\chi^0 &= 1 + (adu)(-cd\theta \sin \theta \cos \varphi + bd\theta \sin \theta \sin \varphi + \cos \theta) \cong 1 + adu \cos \theta, \\ \chi^1 &= \sin \theta \cos \varphi + cd\theta \cos \theta, \quad \chi^2 = \sin \theta \sin \varphi - bd\theta \cos \theta, \\ \chi^3 &= (-cd\theta \sin \theta \cos \varphi + bd\theta \sin \theta \sin \varphi + \cos \theta) - adu.\end{aligned}\tag{8.9.64}$$

Plugging these equations and (8.9.56a) into (8.9.57) yields (written as a row matrix to save space)

$$\dot{k}^\mu = (a \cos \theta, c \cos \theta, -b \cos \theta, -c \sin \theta \cos \varphi + b \sin \theta \sin \varphi + a).\tag{8.9.65}$$

Finally, plugging (8.9.56b), (8.9.56c) and the above equation into (8.9.55) we find all the components of η_{ab} in the system $\{u, r, \theta, \varphi\}$, and then plugging the results into (8.9.53) yields all the components of the Kinnersley metric g_{ab} in $\{u, r, \theta, \varphi\}$. The result will be (8.9.39), which essentially agrees with (13) and (14) of Kinnersley (1969) up to some sign differences. The sign differences come from two reasons: ① The signature in this paper is different from ours; ② the a and c we defined correspond to $-a$ and $-c$ in this paper.

[The End of Optional Reading 8.9.2]

8.10 Coordinate Conditions, the Gauge Freedom of General Relativity

8.10.1 Coordinate Conditions

The vacuum Einstein equation

$$G_{ab} = 0\tag{8.10.1}$$

is a tensor equation. To solve it, one can choose a suitable coordinate system and write it as a system of component equations

$$G_{\mu\nu}(x) = 0, \quad \mu, \nu = 0, 1, 2, 3,\tag{8.10.2}$$

where the x in $G_{\mu\nu}(x)$ indicates that each $G_{\mu\nu}$ is a function of 4 coordinates. Since

$$G_{\mu\nu}(x) = R_{\mu\nu}(x) - \frac{1}{2}R(x)g_{\mu\nu}(x),$$

where $R_{\mu\nu}(x)$ and $R(x)$ can be expressed in terms of $g_{\mu\nu}(x)$ and its partial derivatives, (8.10.2) can be viewed as a system of partial differential equations for the unknown functions $g_{\mu\nu}(x)$. Also, since $g_{\mu\nu} = g_{\nu\mu}$, $g_{\mu\nu}(x)$ only contains 10 independent undetermined functions. On the other hand, due to the symmetry of μ and ν , (8.10.2) also contains 10 algebraically independent partial differential equations. Under suitable boundary conditions, it is reasonable that 10 independent equations could determine 10 independent functions. However, things are not as simple as

that. The curvature tensor R_{abc}^d satisfies the Bianchi identity $\nabla_{[a}R_{bc]d}^e = 0$, from which we have $\nabla_a G^a_b = 0$ [Equation (3.4.17)]. Written in terms of components, this corresponds to 4 differential *identities* satisfied by the functions $g_{\mu\nu}(x)$:

$$G^\mu{}_{\nu;\mu} = 0, \quad (8.10.3)$$

and thus there are only $10 - 4 = 6$ independent functions. The 10 undetermined functions $g_{\mu\nu}(x)$ only have to satisfy 6 independent differential equations; is not that too much freedom? In fact it is indeed so. The key point is that (8.10.2) is the component equation system of the tensor equation $G_{ab} = 0$, and the undetermined functions $g_{\mu\nu}(x)$ are the components of the metric tensor g_{ab} ; if the functions $g_{\mu\nu}(x)$ form a solution to the equation system (8.10.2), then the tensor g_{ab} formed by $g_{\mu\nu}(x)$ together with the coordinate basis satisfies the tensor equation $G_{ab} = 0$, and so a new set of functions $g'_{\mu\nu}(x')$ transferred from $g_{\mu\nu}(x)$ based on the tensor components transformation law is also a solution to (8.10.2). In general, $g_{\mu\nu}$ (as a function of x) and $g'_{\mu\nu}(x')$ (as a function of x') have different functional forms, and hence $g_{\mu\nu}$ and $g'_{\mu\nu}(x')$ are two different sets of solutions to (8.10.2). Thus, boundary conditions can only determine the solution of (8.10.2) “up to a coordinate transformation”; that is, it determines a unique spacetime geometry, but it cannot determine which coordinate system should be used. (This is quite reasonable: the choice of the coordinate system is arbitrary, so it would be strange if the coordinate system can be determined). For instance, the Schwarzschild solution (8.3.18) can be specified the following 10 functions $g_{\mu\nu}(x)$:

$$\begin{aligned} g_{00}(r) &= -(1 - 2M/r), & g_{11}(r) &= (1 - 2M/r)^{-1}, & g_{22}(r) &= r^2, & g_{33}(r) &= r^2 \sin^2 \theta, \\ g_{01} &= g_{02} = g_{03} = g_{12} = g_{13} = g_{23} = 0. \end{aligned} \quad (8.10.4)$$

Define a new coordinate system $\{t', r', \theta', \varphi'\}$ (**isotropic coordinate system**) as follows:

$$t = t' \quad r = r'(1 + M/2r')^2, \quad \theta = \theta' \quad \varphi = \varphi', \quad (8.10.5)$$

then (8.3.18) becomes

$$\begin{aligned} ds^2 &= -[(1 - M/2r')/(1 + M/2r')]^2 dt'^2 \\ &\quad + (1 + M/2r')^4 [dr'^2 + r'^2(d\theta'^2 + \sin^2 \theta' d\varphi'^2)], \end{aligned} \quad (8.10.6)$$

and the 10 functions $g'_{\mu\nu}(x')$ representing it are different from those in (8.10.4). For example, the dependence of $g'_{00}(r') = -[(1 - M/2r')/(1 + M/2r')]^2$ on its argument r' is obviously different from the dependence of $g_{00}(r)$ on its argument r . However, all $R'_{\mu\nu}$ derived from $g'_{\mu\nu}(x')$ also vanish, and thus (8.10.6) and (8.3.18) are both (spherically symmetric) solutions to the vacuum Einstein equations $G_{\mu\nu} = 0$ that satisfy the same boundary conditions, and therefore they represent the same

geometry. This is an example of “boundary conditions cannot determine a unique solution, but they can determine a unique geometry”.

Since a coordinate transformation involves 4 arbitrary functions (new coordinates expressed using old coordinates), we can say that general covariance provides 4 “degrees of freedom” for the equation system (8.10.2). To remove this uncertainty, one needs to assign a specific coordinate system, i.e., assign 4 additional equations for the functions $g_{\mu\nu}(x)$; these 4 equations together are called a **coordinate condition**. The 4 equations below are an example of a coordinate condition:

$$g_{00} = -1, \quad g_{0i} = 0 \quad (i = 1, 2, 3). \quad (8.10.7)$$

The coordinates satisfying this condition are called Gaussian normal coordinates (see Optional Reading 8.10.1 for details). Another example of a coordinate condition is requiring the coordinates x^σ to satisfy the following 4 equations:

$$g^{ab} \nabla_a \nabla_b x^\sigma = 0 \quad (\sigma = 0, 1, 2, 3). \quad (8.10.8)$$

Calculation shows that [see Weinberg (1972) pp. 161–163] the above equations are equivalent to the following 4 equations:

$$g^{\mu\nu} \Gamma^\lambda_{\mu\nu} = 0 \quad (\lambda = 0, 1, 2, 3). \quad (8.10.8')$$

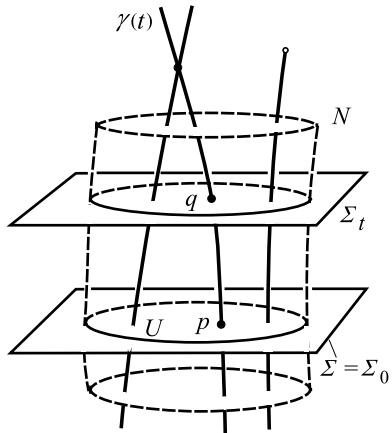
Equation (8.10.8) or (8.10.8') is called the **harmonic coordinate condition**, since a function f satisfying $g^{ab} \nabla_a \nabla_b f = 0$ is called a harmonic function. Equation (8.10.8') indicates more clearly that this coordinate condition is indeed additional equations restricting the functions $g_{\mu\nu}(x)$.

A coordinate condition is obviously not a generally covariant equation, since its mission is to pick a special coordinate system to remove the uncertainty of $g_{\mu\nu}(x)$ coming from the general covariance of Einstein's equation. A coordinate condition should also satisfy the following requirement: starting from an arbitrary set of functions $g_{\mu\nu}(x)$, one can always find $g'_{\mu\nu}(x')$ satisfying the coordinate condition by a coordinate transformation.

[Optional Reading 8.10.1]

A Gaussian normal coordinate system is a coordinate system defined by means of geodesics. Suppose Σ is an arbitrary spacelike hypersurface in a spacetime (M, g_{ab}) , n^a is a unit normal vector field on Σ , and $\{x^i\}$ is an arbitrary 3-dimensional coordinate system on an open region $U \subset \Sigma$. Any point p in U together with its unit normal vector $n^a|_p$ (normal to Σ) determines a unique geodesic $\gamma(t)$ [stipulate $t(p) = 0$], which is orthogonal to Σ . Although these geodesics emanating from U may intersect (see Fig. 8.16) or run into other unideal situations, it can be proved that as long as we take an appropriate U , there must be an open subset $N \subset M$ containing U , in which for any point q there exists a unique p in U such that q is on the geodesic $\gamma(t)$ staring from p . Define $x^i|_q \equiv x^i|_p$ and take the value $t(q)$ of the parameter of $\gamma(t)$ at q as the zeroth coordinate, then the coordinate system $\{t, x^i\}$ (with N as the coordinate patch) is called a **Gaussian normal coordinate system**. Now we will show that Gaussian normal coordinates satisfy (8.10.7). Noticing that the geodesic $\gamma(t)$ is a t -coordinate line, whose tangent $(\partial/\partial t)^a$ is the zeroth coordinate basis vector, we have

Fig. 8.16 Constructing a Gaussian normal coordinate system from Σ , whose coordinate patch is N



$$g_{00} = g_{ab} (\partial/\partial t)^a (\partial/\partial t)^b.$$

Also $(\partial/\partial t)^a|_p = n^a|_p$, and hence $g_{00}|_p = (n_a n^a)|_p = -1$. Since the tangent vector $(\partial/\partial t)^a$ is transported parallelly along $\gamma(t)$, and parallel transport preserves the inner product, we have

$$g_{00}|_q = [(\partial/\partial t)_a (\partial/\partial t)^a]|_q = [(\partial/\partial t)_a (\partial/\partial t)^a]|_p = -1, \quad \forall q \in N. \quad (8.10.9)$$

Let Σ_t be a constant- t hypersurface, then the x^i -coordinate lines are lying on Σ_t since their t are constant, and hence the three spatial coordinate basis vectors $(\partial/\partial x^i)^a$ are everywhere tangent to Σ_t . Since $g_{0i} = g_{ab} (\partial/\partial t)^a (\partial/\partial x^i)^b$, in order to prove $g_{0i} = 0$ ($i = 1, 2, 3$) we only have to show that $\gamma(t)$ is orthogonal to any Σ_t . From the construction of $\gamma(t)$ we can see that this is surely correct for $\Sigma_0 \equiv \Sigma$, i.e., $g_{0i}|_p = 0$, and therefore we only have to show that $g_{0i} = (\partial/\partial t)_a (\partial/\partial x^i)^a$ is a constant along $\gamma(t)$. The following derivation indicates that this is indeed true:

$$\begin{aligned} (\partial/\partial t)^b \nabla_b [(\partial/\partial t)_a (\partial/\partial x^i)^a] &= (\partial/\partial t)_a (\partial/\partial t)^b \nabla_b (\partial/\partial x^i)^a = (\partial/\partial t)_a (\partial/\partial x^i)^b \nabla_b (\partial/\partial t)^a \\ &= \frac{1}{2} (\partial/\partial x^i)^b \nabla_b [(\partial/\partial t)_a (\partial/\partial t)^a] = \frac{1}{2} (\partial/\partial x^i)^b \nabla_b g_{00} = 0, \end{aligned}$$

where the geodesic equation is used in the first equality, the commutativity of the coordinate basis vectors is used in the second equality, the Leibniz rule is used in the third equality, and (8.10.9) is used in the last step.

[The End of Optional Reading 8.10.1]

Now we discuss the Einstein equations with source $G_{\mu\nu} = 8\pi T_{\mu\nu}$. Suppose the matter field has N components, then usually it needs to satisfy N equations (such as equations of motion). If the equations are independent (see Example 2 for the non-independent case), then combining them with the Einstein equations we obtain $10 + N$ equations. It seems that they can determine $10 + N$ functions. However, the $10 g_{\mu\nu}$ automatically satisfy $G^{\mu}_{\nu;\mu} = 0$, and the equations of motion of the matter field automatically lead to $T^{\mu}_{\nu;\mu} = 0$, and hence $G^{\mu}_{\nu;\mu} - 8\pi T^{\mu}_{\nu;\mu}$ automatically vanishes. That is, we have the following differential identities

$$G^\mu_{\nu;\mu} - 8\pi T^\mu_{\nu;\mu} = 0, \quad \nu = 0, 1, 2, 3, \quad (8.10.10)$$

which will “dispose of” 4 equations. Together with 4 coordinate conditions, these equations determine $10 + N$ unknown functions exactly.

Example 1 Suppose the matter field is a perfect fluid, whose components contain the proper density ρ , pressure p and the 4-velocity components U^μ , and hence $N = 6$. The equations they satisfy are: (a) the equation of state $f(\rho, p) = 0$, where f is a certain function [see above (9.3.20)], (b) the divergence-free condition $\nabla^a T_{ab} = 0$ for the energy-momentum tensor,¹⁶ (c) the normalization conditions $g_{\mu\nu} U^\mu U^\nu = -1$ for the 4-velocity. In total there are $1 + 4 + 1 = 6$ equations, which agrees with the generic discussion above.

Example 2 Suppose the matter field is a source-free electromagnetic field. The 4-potential only has to satisfy the equation of motion

$$\nabla^a \nabla_a A_b - \nabla^a \nabla_b A_a = 0, \quad [\text{a special case of (7.2.7)}] \quad (8.10.11)$$

and thus the numbers of the component equations and field components are both 4. However, among the 4 equations above only 3 are independent, since A_a (or any 1-form) satisfies the following differential identity (which can be proved following the proof of Exercise 7.1)

$$\nabla^b \nabla^a (\nabla_a A_b - \nabla_b A_a) = 0, \quad (\text{i.e., } \nabla^b \nabla^a F_{ab} = 0) \quad (8.10.12)$$

which will “dispose of” one equation, and make (8.10.11) one equation short. This is caused by the gauge freedom of A_a , and so after adding the Lorenz condition $\nabla^a A_a = 0$ (choosing a gauge), we can apply the generic discussion above. Assigning a gauge condition here is similar to assigning a coordinate condition for $g_{\mu\nu}$. As a matter of fact, the latter is also some kind of gauge choice, see the next subsection for details.

Finally, we should point out that for partial differential equations, the claim “given suitable boundary conditions, there is a unique solution as long as the number of equations is equal to the number of the undetermined functions” is not as simple as that for ordinary differential equations. There are many subtleties in this case. One may view this subsection as a hand-waving discussion (for illustrating the necessity of coordinate conditions), and should not regard it as a rigorous analysis.

¹⁶ From Sect. 6.5 we can see that the divergence-free condition $\partial^a T_{ab} = 0$ for the energy-momentum tensor of a perfect fluid in Minkowski spacetime contains the equations of motion of the fluid, namely (6.5.7) and (6.5.8), which has in total $1 + 3 = 4$ equations. For a curved spacetime, the condition $\nabla^a T_{ab} = 0$ also leads to 4 similar equations.

8.10.2 The Gauge Freedom of General Relativity

The above discussion can also be formulated using the geometric language, i.e., instead of talking about the component equations $G_{\mu\nu} = 8\pi T_{\mu\nu}$, we can discuss the tensor equation $G_{ab} = 8\pi T_{ab}$. Take the vacuum field equation $G_{ab} = 0$ as an example. One can prove the following claim (see later): suppose $\phi : M \rightarrow M$ is a diffeomorphism, $R_{ab}[g]$ is the Ricci tensor of the metric g_{ab} , then

$$\phi_*(R_{ab}[g]) = R_{ab}[\phi_*g]. \quad (8.10.13)$$

From this we can easily get $G_{ab}[g] = 0 \Leftrightarrow G_{ab}[\phi_*g] = 0$. This indicates that g_{ab} is a solution to $G_{ab} = 0$ if and only if ϕ_*g_{ab} is also a solution. Thus, the boundary conditions can only determine a solution g_{ab} to Einstein's equation up to a diffeomorphism. This is actually the active formulation (see Optional Reading 4.1.1) equivalent to the passive version above that “boundary conditions can only determine $g_{\mu\nu}$ up to a coordinate transformation”. In the passive formulation, the components $g_{\mu\nu}$ and $g'_{\mu\nu}$ of the same metric field g_{ab} in different coordinate systems represent the same (local) geometry; in the active formulation, suppose $\phi : M \rightarrow \tilde{M}$ is a diffeomorphism, then g_{ab} and $\tilde{g}_{ab} \equiv \phi_*g_{ab}$ represent the same geometry. In order to avoid confusion, first we consider two manifolds M and \tilde{M} . If there exists a diffeomorphism $\phi : M \rightarrow \tilde{M}$, then M and \tilde{M} “cannot be more alike”. Then, we consider two spacetimes (or more generally, two generalized Riemannian spaces) (M, g_{ab}) and $(\tilde{M}, \tilde{g}_{ab})$. If there exists a diffeomorphism $\phi : M \rightarrow \tilde{M}$ and $\phi_*g_{ab} = \tilde{g}_{ab}$, then these two spacetimes “cannot be more alike”, i.e., they have the same spacetime geometry, and every phenomenon that can be described by (M, g_{ab}) can be described equivalently by $(\tilde{M}, \tilde{g}_{ab})$. For instance, suppose there are two vectors u^a and v^b at a point p in M , then there are two corresponding vectors ϕ_*u^a and ϕ_*v^b at the point $\phi(p)$ in \tilde{M} . In addition, the inner product of ϕ_*u^a and ϕ_*v^b , $\tilde{g}_{ab}|_{\phi(p)}(\phi_*u)^a(\phi_*v)^b$, equals the inner product of u^a and v^b , $g_{ab}|_p u^a v^b$, because

$$g_{ab}|_p u^a v^b = (\phi^* \tilde{g})_{ab}|_p u^a v^b = \tilde{g}_{ab}|_{\phi(p)}(\phi_*u)^a(\phi_*v)^b.$$

One can also show that the tensor product of ϕ_*u^a and ϕ_*v^b corresponds to the tensor product of u^a and v^b , i.e., $(\phi_*u^a)(\phi_*v^b) = \phi_*(u^a v^b)$, etc. In short, we have at $\phi(p)$ whatever we have at p , and we can do at $\phi(p)$ whatever we can do at p and get the same result (matched by ϕ_*). If we consider the metric at p as a stage, and consider manipulating the quantities at p as putting on a play, one can say colloquially that ϕ_* “carries the stage” of (M, g_{ab}) to $(\tilde{M}, \tilde{g}_{ab})$ so that we can “perform a play in a different town” (i.e., manipulate the pushforward of the quantities at a different point).

This discussion can also be applied to the case where $M = \tilde{M}$. Suppose on M we have a metric field g_{ab} and a diffeomorphism $\phi : M \rightarrow M$, then based on the discussion that (M, g_{ab}) and $(\tilde{M}, \tilde{g}_{ab})$ “cannot be more alike”, we can see that (M, g_{ab}) and (M, ϕ_*g_{ab}) are equivalent geometrically. However, one should notice

that now there are two metrics $g_{ab}|_p$ and $\phi_*g_{ab}|_p$ at a point p in M . Suppose u^a and v^a are vectors at p , by (M, g_{ab}) and (M, ϕ_*g_{ab}) are equivalent we do not mean that $g_{ab}|_p u^a v^b = (\phi_*g)_{ab}|_p u^a v^b$ (this only holds when ϕ is an isometry), instead we mean that $g_{ab}|_p u^a v^b = (\phi_*g)_{ab}|_{\phi(p)}(\phi_*u)^a(\phi_*v)^b$, i.e., we can “carry the whole stage and perform the same play at $\phi(p)$ ”. Here we give an application example. Let R_{abc}^d and \tilde{R}_{abc}^d represent the Riemann tensor fields of g_{ab} and $\tilde{g}_{ab} \equiv \phi_*g_{ab}$, respectively. Given $R_{abc}^d|_p$ we would like to find $\tilde{R}_{abc}^d|_{\phi(p)}$. Knowing that we can “perform the same play in a different town”, all we have to do is to push forward $R_{abc}^d|_p$ to $\phi(p)$ using ϕ_* . More precisely speaking, when calculating $R_{abc}^d|_p$ we have done the following manipulation: first find the ∇_a associated with g_{ab} , then find $R_{abc}^d|_p$ from $(\nabla_a \nabla_b - \nabla_b \nabla_a)\omega_c = R_{abc}^d \omega_d$. This manipulation is just like “performing a play”. In order to find $\tilde{R}_{abc}^d|_{\phi(p)}$, in principle we need to perform a similar manipulation: first find the $\tilde{\nabla}_a$ associated with \tilde{g}_{ab} , then find $\tilde{R}_{abc}^d|_p$ from $(\tilde{\nabla}_a \tilde{\nabla}_b - \tilde{\nabla}_b \tilde{\nabla}_a)\omega_c = \tilde{R}_{abc}^d \omega_d$. Nevertheless, it is in fact not necessary to do it all over again like this, because it is natural to believe that as long as we push forward the result of the manipulation on g_{ab} (and quantities derivable from it) at p to $\phi(p)$ using ϕ_* , it must be equal to the result of the manipulation on \tilde{g}_{ab} (and quantities derivable from it) at $\phi(p)$. That is, we can believe that

$$\phi_*(R_{abc}^d|_p) = \tilde{R}_{abc}^d|_{\phi(p)}. \quad (8.10.14)$$

For all quantities determined by g_{ab} (all geometric quantities), such as R_{ab} , R , G_{ab} , etc., we have similar relations, and thus (8.10.13) holds. If you want, the reader can also verify (8.10.14) by computing it directly; hint: first verify that the $\tilde{\nabla}_a$ associated with \tilde{g}_{ab} satisfies

$$\tilde{\nabla}_a(\phi_*T) = \phi_*(\nabla_a T) \quad (\text{where } T \text{ is a tensor field of any type}). \quad (8.10.15)$$

In a word, in the sense of “performing the same play in a different town” we can say that two metric fields g_{ab} and ϕ_*g_{ab} on M (when ϕ is a diffeomorphism) describe the same geometry, or say that g_{ab} and ϕ_*g_{ab} are equivalent. Thus, metric fields do not have a one-to-one correspondence with spacetime geometries; instead, one kind of spacetime geometry corresponds to an equivalent class $\{g_{ab}\}$. This is similar to the fact that in the theory of electromagnetism a gauge transformation of the 4-potential A_a does not change the electromagnetic field F_{ab} . Therefore, the property that “changing g_{ab} to ϕ_*g_{ab} does not change the geometry” is called the **gauge freedom** of general relativity. As a physical theory, general relativity is endowed with gauge freedom, which is an important feature of this theory (just like the electromagnetic theory formulated by the 4-potential is endowed with gauge freedom). This gauge freedom has great significance for further studying relativity. For instance, it will play an important role in Chaps. 14 and 16. The concepts of “gauge transformation” and “gauge invariance” came originally from the theory of electromagnetism, and have become extremely important in theoretical physics. Roughly speaking, any transformation that does not change the essence of the physics can be called

a gauge transformation, and the corresponding invariance (freedom) is then called the gauge invariance (freedom). For convenience of computation, when discussing a specific problem one can choose a certain gauge, which is called “gauge fixing”. As for general relativity, choosing a coordinate system is nothing but fixing a gauge. Besides the transformation of the electromagnetic 4-potential, there are mainly two other kinds of gauge transformation in this text so far: ① the gauge transformation in the theory of linearized gravity [Equation (7.8.14)], ② the gauge transformation in general relativity. (In the active language this is a diffeomorphism $\phi : M \rightarrow M$, and in the passive language this is a coordinate transformation). Actually, ① is just an infinitesimal version of ②, and the reason is as follows. The gauge transformation in the theory of linearized gravity is given by

$$\gamma_{ab} \mapsto \tilde{\gamma}_{ab} = \gamma_{ab} + \partial_a \xi_b + \partial_b \xi_a$$

(where ξ^a is an “infinitesimal” vector field). The difference between the metric before and after the transformation, namely $g_{ab} = \eta_{ab} + \gamma_{ab}$ and $\tilde{g}_{ab} = \tilde{\eta}_{ab} + \tilde{\gamma}_{ab}$, is

$$\tilde{g}_{ab} - g_{ab} = \partial_a \xi_b + \partial_b \xi_a. \quad (8.10.16)$$

Introduce a vector λ^a and a real number t to express ξ^a as $\xi^a = t\lambda^a$ (where t is a small quantity of the same order as ξ^a , i.e., a first-order small quantity), then it follows from the formula of the Lie derivative that $\mathcal{L}_\lambda \eta_{ab} = \partial_a \lambda_b + \partial_b \lambda_a$. Since

$$\mathcal{L}_\lambda \eta_{ab} = \mathcal{L}_\lambda(g_{ab} - \gamma_{ab}) \cong \mathcal{L}_\lambda g_{ab}$$

(the second term is ignored as it is a second-order small term), we have

$$\partial_a \lambda_b + \partial_b \lambda_a \cong \mathcal{L}_\lambda g_{ab} \cong \frac{\phi_t^* g_{ab} - g_{ab}}{t}. \quad (8.10.17)$$

Comparing this with (8.10.16) yields $\phi_t^* g_{ab} - g_{ab} \cong \tilde{g}_{ab} - g_{ab}$, and hence $\tilde{g}_{ab} \cong \phi_t^* g_{ab} - g_{ab}$. Thus, the original metric g_{ab} and the new metric \tilde{g}_{ab} after the transformation only differ by a diffeomorphism under the first-order approximation.

Of course, it is not just the spacetime geometry that we care about, but also physics. Here is a general conclusion: suppose a physical theory is described by a manifold M and some tensor fields $T^{(i)}$ living on it (for instance, for an electrovac spacetime, $T^{(i)}$ includes at least g_{ab} and F_{ab}), then $(M, T^{(i)})$ and $(M, \tilde{T}^{(i)})$ describe the same physics if and only if there exists a diffeomorphism $\phi : M \rightarrow M$ such that $\tilde{T}^{(i)} = \phi_* T^{(i)}$.

Exercises

- ~8.1. Prove Proposition 8.1.1.
- ~8.2. Suppose $\gamma(r)$ is a curve from p_1 to p_2 on Σ_t in Fig. 8.7 where θ and φ are both constants (with the radial coordinate r as the parameter of the curve). Show that $\gamma(r)$ is a (non-affinely parametrized) geodesic. Hint: use (5.7.2).
- ~8.3. Suppose ξ^a is a timelike Killing vector field in a stationary spacetime, and $\chi \equiv \sqrt{-g_{ab}\xi^a\xi^b}$.
- Show that χ is a constant along an integral curve of ξ^a ;
 - Show that the 4-acceleration $A^a = \nabla^a(\ln \chi)$. Hint: use the Killing equation $\nabla^{(a}\xi^{b)} = 0$ and the result of (a).
- ~8.4. Show that: (a) the trace of the energy-momentum tensor of an electromagnetic field is zero, i.e., $T \equiv g^{ab}T_{ab} = 0$; (b) the scalar curvature of an electrovac spacetime is $R = 0$.
- ~8.5. Prove (8.4.7) and (8.4.28).
- 8.6. Suppose F_{ab} is a 2-form field in an arbitrary spacetime, ${}^*F_{ab}$ is the dual 2-form field of F_{ab} , and $\alpha \in [0, 2\pi]$ is a constant real number, then $F'_{ab} \equiv F_{ab} \cos \alpha - {}^*F_{ab} \sin \alpha$ is called a **duality rotation** of F_{ab} with the angle α .
- Show that F_{ab} is a source-free electromagnetic field if and only if F'_{ab} is a source-free electromagnetic field. [The proof is straightforward. One can see this directly from the exterior differential expressions (7.2.4') and (7.2.5') of Maxwell's equations].
 - Show that the electromagnetic fields F_{ab} and F'_{ab} have the same energy-momentum tensor. Hint: the proof can be simplified by using the symmetric expression (6.6.28').
 - Let $M \equiv 2F_{ab}F^{ab}$, $N \equiv 2F_{ab}{}^*F^{ab}$, $M' \equiv 2F'_{ab}F'^{ab}$, $N' \equiv 2F'_{ab}{}^*F'^{ab}$. Show that

$$M' = M \cos 2\alpha - N \sin 2\alpha, \quad N' = M \sin 2\alpha + N \cos 2\alpha.$$

(d) Let $\Sigma_{ab} \equiv F_{ab} + i{}^*F_{ab}$, and $\Sigma'_{ab} \equiv F'_{ab} + i{}^*F'_{ab}$, then $K \equiv \Sigma_{ab}\Sigma^{ab}$ and $K' \equiv \Sigma'_{ab}\Sigma'^{ab}$ are complex scalar fields, and hence the K and K' at each spacetime point correspond to two vectors in the complex plane. Using the result of (c) show that the vector K' is the result of rotating the vector K counterclockwise by an angle 2α (i.e., $|K| = |K'|$, and the arguments of K' and K differ by 2α).

(e) Suppose (\vec{E}, \vec{B}) and (\vec{E}', \vec{B}') are the electric and magnetic fields of F_{ab} and F'_{ab} measured by an instantaneous observer, respectively. Show that

$$\vec{E}' = \vec{E} \cos \alpha + \vec{B} \sin \alpha, \quad \vec{B}' = -\vec{E} \sin \alpha + \vec{B} \cos \alpha. \quad (8.10.18)$$

NB: For further interpretations of the physical meaning of the dual rotation, see Volume II and Jackson (1998).

- 8.7. An n -dimensional spacetime is called an **Einstein spacetime** if $R_{ab} = Rg_{ab}/n$, where g_{ab} , R_{ab} and R are the metric, Ricci tensor and scalar curvature, respectively. Show that an electrovac spacetime (where the electromagnetic field is not vanishing) is not an Einstein spacetime. NB: It follows from Exercise 3.17 that any 2-dimensional spacetime must be an Einstein spacetime.
- 8.8. Consider Taub's plane symmetric vacuum solution (8.6.1').
 (a) Write down the expression for the 4-velocity of a static observer in terms of the coordinate basis vectors; (b) Suppose the spatial coordinates of two static observers are (x, y, z_1) and (x, y, z_2) , respectively. Find the spatial distance between them.
- 8.9. Show that the F_{ab} in (8.6.5) has plane symmetry, i.e., $\mathcal{L}_{\xi_i} F_{ab} = 0$ ($i = 1, 2, 3$), where $\xi_1^a \equiv (\partial/\partial x)^a$, $\xi_2^a \equiv (\partial/\partial y)^a$, $\xi_3^a \equiv -y(\partial/\partial x)^a + x(\partial/\partial y)^a$ are the Killing fields reflecting the plane symmetry of the metric (8.6.3).
- *8.10. Derive the expressions for the Maxwell equations with source in the NP formalism. Answer: For each of the equations in (8.8.3), one needs to add a term to the right-hand side. In sequence they are $-4\pi J_4$, $-4\pi J_2$, $-4\pi J_1$, $-4\pi J_3$ (where J_1, J_2, J_3, J_4 are the components of J_a in the null tetrad).
- *8.11. Prove (8.8.7) and (8.8.10).

References

- Bonnor, W. B. (1994), ‘The photon rocket’, *Class. Quant. Grav.* **11**, 2007–2012.
- Carmeli, M. (1982), *Classical Fields General Relativity and Gauge Theory*, John Wiley & Sons, New York.
- Damour, T. (1995), ‘Photon rockets and gravitational radiation’, *Class. Quant. Grav.* **12**, 725–738. [arXiv:gr-qc/9412063](https://arxiv.org/abs/gr-qc/9412063).
- Dain, S., Moreschi, O. M. and Gleiser, R. J. (1996), ‘Photon rockets and the Robinson-Trautman geometries’, *Class. Quant. Grav.* **13**, 1155–1160. [arXiv:gr-qc/0203064](https://arxiv.org/abs/gr-qc/0203064).
- Hawking, S. W. and Ellis, G. F. R. (1973), *The Large Scale Structure of Space-Time*, Cambridge University Press, Cambridge.
- Jackson, J. D. (1998), *Classical Electrodynamics*, John Wiley & Sons, Inc., New York.
- Kinnersley, W. (1969), ‘Field of an arbitrarily accelerating point mass’, *Phys. Rev.* **186**, 1335–1336.
- Kuang, Z. and Liang, C. (1988), ‘Birkhoff and Taub theorems generalized to metrics with conformal symmetries’, *J. Math. Phys.* **29**, 2475–2478.
- Kuang, Z., Li, J. and Liang, C. (1986), ‘Gauge freedom of plane-symmetric line elements with semi-plane-symmetric null electromagnetic fields’, *Phys. Rev. D* **34**, 2241–2245.
- Kuang, Z., Li, J. and Liang, C. (1987), ‘Completion of plane-symmetric metrics yielded by electromagnetic fields’, *Gen. Rela. Grav.* **19**, 345–350.
- Letelier, P. S. and Tabenski, R. R. (1974), ‘The general solution to Einstein-Maxwell equations with plane symmetry’, *J. Math. Phys.* **15**, 594.
- Li, J. and Liang, C. (1985), ‘An extension of the plane-symmetric electrovac general solution to Einstein equations’, *Gen. Rela. Grav.* **17**, 1001–1013.
- Li, J. and Liang, C. (1989), ‘Static semi-plane-symmetric metrics yielded by plane-symmetric electromagnetic fields’, *J. Math. Phys.* **30**, 2915–2917.

- Liang, C. (1995), 'A family of cylindrically symmetric solutions to Einstein-Maxwell equations', *Gen. Rela. Grav.* **27**, 669–677.
- Newman, E. and Penrose, R. (1962), 'An approach to gravitational radiation by a method of spin coefficients', *J. Math. Phys.* **3**, 566.
- Patnaik, S. (1970), 'Einstein-Maxwell fields with plane symmetry', *Proc. Camb. Phil. Soc.* **67**, 127.
- Stephani, H. (1982), *General Relativity: An Introduction to the Theory of Gravitational Field*, Cambridge University Press, Cambridge.
- Stephani, H., Kramer, D., MacCallum, M. A. H., Hoenselaers, C. and Herlt, E. (2003), *Exact Solutions of Einstein's Field Equations*, Cambridge University Press, Cambridge.
- Tariq, N. and Tupper, B. O. J. (1976), 'Einstein-Maxwell metrics admitting a dual interpretation', *J. Math. Phys.* **17**, 292–296.
- Taub, H. (1951), 'Empty space-times admitting a three parameter group of motions', *Ann. Math.* **53**, 472.
- Wald, R. M. (1984), *General Relativity*, The University of Chicago Press, Chicago.
- Weinberg, S. (1972), *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*, John Wiley and Sons, New York.

Chapter 9

Schwarzschild Spacetimes



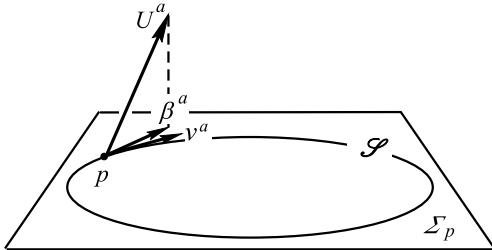
In the first three sections of Chap. 8 we had a discussion on static spherically symmetric metrics and the vacuum Schwarzschild solution, which focused mainly on finding the solution. In view of the essentialness of the Schwarzschild solution, this chapter will further discuss several intimately related problems: Sect. 9.1 discusses the timelike and null geodesics in Schwarzschild spacetime; Sect. 9.2 introduces three experimental tests of general relativity posed by Einstein using the vacuum Schwarzschild solution in his early years, namely the gravitational redshift, the precession of the perihelion of Mercury and the bending of starlight in the Sun's gravitational field; Sect. 9.3 discusses the spacetime geometric structure and physical states in the interior of a spherically symmetric star, as well as the evolution of a spherically symmetric star; Sect. 9.4 analyzes the theory of the extension of the Schwarzschild spacetime in detail.

9.1 Geodesics in Schwarzschild Spacetimes

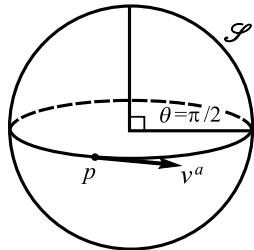
Let $\gamma(\tau)$ be a timelike (or null) geodesic. For a timelike geodesic, τ represents the proper time; for a null geodesic, τ represents a chosen affine parameter. In order to find the parametric equations $x^\mu(\tau)$ of $\gamma(\tau)$, generally we need to solve the following differential equations:

$$\frac{d^2x^\mu}{d\tau^2} + \Gamma^\mu_{\nu\sigma} \frac{dx^\nu}{d\tau} \frac{dx^\sigma}{d\tau} = 0, \quad \mu = 0, 1, 2, 3. \quad (9.1.1)$$

Since in these equations the unknown functions $x^\mu(\tau)$ and their derivatives are coupled with each other, solving for them is in general not simple. However, if the spacetime has a sufficient amount of Killing vector fields, one can find $x^\mu(\tau)$ in a clever way using Theorem 4.3.3. Schwarzschild spacetime is an example of this.



(a) β^a is the projection of the 4-velocity U^a of geodesic $\gamma(\tau)$ at a point p . v^a is the component of β^a tangent to the orbit sphere.



(b) A point p and the vector v^a on the orbit sphere S determine a unique geodesic (a great circle).

Fig. 9.1 Figures for the proof of Proposition 9.1.1

Before applying this theorem, we can also simplify the coordinate representation of the geodesic $\gamma(\tau)$ using the spherical symmetry of Schwarzschild spacetime.

Proposition 9.1.1 *Suppose $\gamma(\tau)$ is a timelike or null geodesic in Schwarzschild spacetime, then one can always choose the Schwarzschild coordinates such that $\theta = \pi/2$ along $\gamma(\tau)$, in other words, such that $\gamma(\tau)$ lies in the “equatorial plane”.*

Proof $\forall p \in \gamma(\tau)$, the orbit sphere S passing through p (see Definition 1 in Sect. 8.2) must lie in the constant- t surface Σ_t passing through p (Fig. 9.1a). The choice of the coordinates θ and φ of the Schwarzschild system is quite arbitrary. Noticing that the Schwarzschild line element (8.3.18) being invariant under the transformation $\theta \rightarrow \pi - \theta$ assures that the northern and southern hemisphere are symmetric with respect to the equator, if one can choose θ and φ such that the value of θ at a point $p \in \gamma(\tau)$ is $\pi/2$ and the θ -component of the 4-velocity $U^a \equiv (\partial/\partial\tau)^a$ vanishes at p , then we will see that the whole $\gamma(\tau)$ has $\theta(\tau) = \pi/2$. Thus, we only have to show that this kind of choice of θ and φ is indeed available. Since the geodesic is not always orthogonal to the constant- t surface (otherwise it will become the world line of a static observer, which is not a geodesic), one can always find a point p whose U^a has a projection $\beta^a \neq 0$ on Σ_p . The orbit surface passing through p is $S \subset \Sigma_p$. If β^a has a component v^a tangent to S , then (p, v^a) determines a unique geodesic on S , i.e., a great circle (see Fig. 9.1b), and we can define the coordinates θ and φ on S with this great circle as the equator; if β^a does not have a component v^a tangent to S [i.e., $\gamma(\tau)$ is a radial geodesic], then we can choose any great circle passing through p as the equator. Using the “carry method” in Sect. 8.3 to carry the θ and φ coordinates away from S , we obtain the Schwarzschild coordinate system $\{t, r, \theta, \varphi\}$ that satisfies our requirements, namely ① $\theta(p) = \pi/2$; ② the components $d\theta/d\tau|_p$ of $(\partial/\partial\tau)^a|_p$ in the coordinate basis $(\partial/\partial\theta)^a|_p$ vanishes. \square

The conclusion that a point $p \in \gamma(\tau)$ satisfying ① and ② assures that $\theta = \pi/2$ along the whole curve can also be proved analytically as follows: from the expression (8.3.20) for the $\Gamma^\sigma_{\mu\nu}$ of the Schwarzschild line element we can see that for $\mu = 2$, (9.1.1) is

$$\frac{d^2\theta}{d\tau^2} + \frac{2}{r} \frac{dr}{d\tau} \frac{d\theta}{d\tau} - \sin\theta \cos\theta \left(\frac{d\varphi}{d\tau} \right)^2 = 0. \quad (9.1.2)$$

Since the geodesic $\gamma(\tau)$ has been given beforehand, the functions $t(\tau)$, $r(\tau)$, $\theta(\tau)$, $\varphi(\tau)$ are all determined once a coordinate system is chosen. In order to show that $\theta = \pi/2$ for the entire curve, we only have to notice that (9.1.2) is a 2nd-order ordinary differential equation, and $\theta(\tau) = \pi/2$ is the unique solution satisfying the initial conditions $\theta(p) = \pi/2$ and $d\theta/d\tau|_p = 0$.

According to Proposition 9.1.1, one can always choose the Schwarzschild coordinates so that the parametric equations for the above-mentioned geodesic $\gamma(\tau)$ are

$$t = t(\tau), \quad r = r(\tau), \quad \theta = \pi/2, \quad \varphi = \varphi(\tau).$$

Suppose $U^a \equiv (\partial/\partial\tau)^a$ is the tangent of $\gamma(\tau)$, and define $\kappa := -g_{ab}U^aU^b$, then

$$\kappa = \begin{cases} 1, & \text{(for timelike geodesics)} \\ 0, & \text{(for null geodesics)} \end{cases},$$

and

$$\begin{aligned} -\kappa &= g_{ab} \left(\frac{\partial}{\partial\tau} \right)^a \left(\frac{\partial}{\partial\tau} \right)^b = g_{00} \left(\frac{dt}{d\tau} \right)^2 + g_{11} \left(\frac{dr}{d\tau} \right)^2 + g_{22} \left(\frac{d\theta}{d\tau} \right)^2 + g_{33} \left(\frac{d\varphi}{d\tau} \right)^2 \\ &= - \left(1 - \frac{2M}{r} \right) \left(\frac{dt}{d\tau} \right)^2 + \left(1 - \frac{2M}{r} \right)^{-1} \left(\frac{dr}{d\tau} \right)^2 + r^2 \left(\frac{d\varphi}{d\tau} \right)^2, \end{aligned} \quad (9.1.3)$$

where in the last step we used $\theta = \pi/2$. Noticing that $(\partial/\partial t)^a$ and $(\partial/\partial\varphi)^a$ are Killing vector fields, by means of Theorem 4.3.3, we can define two constants on the geodesic $\gamma(\tau)$:

$$E := -g_{ab} \left(\frac{\partial}{\partial t} \right)^a \left(\frac{\partial}{\partial\tau} \right)^b = -g_{00} \frac{dt}{d\tau} = \left(1 - \frac{2M}{r} \right) \frac{dt}{d\tau}, \quad (9.1.4)$$

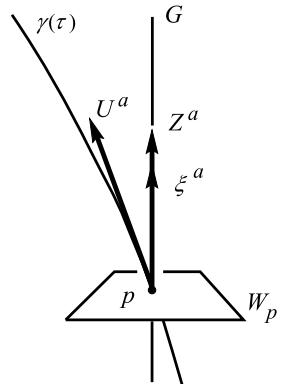
$$L := g_{ab} \left(\frac{\partial}{\partial\varphi} \right)^a \left(\frac{\partial}{\partial\tau} \right)^b = g_{33} \frac{d\varphi}{d\tau} = r^2 \frac{d\varphi}{d\tau}, \quad (9.1.5)$$

Plugging (9.1.4) and (9.1.5) into (9.1.3) yields

$$-\kappa = - \left(1 - \frac{2M}{r} \right)^{-1} E^2 + \left(1 - \frac{2M}{r} \right)^{-1} \left(\frac{dr}{d\tau} \right)^2 + \frac{L^2}{r^2}. \quad (9.1.6)$$

This equation, which contains only the unknown function $r(\tau)$ and its 1st-order derivative, is solvable in principle. Plugging the $r(\tau)$ we just obtained into (9.1.4) and (9.1.5), in principle we can find the unknown functions $t(\tau)$ and $\varphi(\tau)$, and hence find the parametric equations of $\gamma(\tau)$.

Fig. 9.2 Local measurement on a freely falling particle $\gamma(\tau)$ made by a static observer G



We now discuss the physical meaning of these two constants E and L . Suppose $\gamma(\tau)$ is a timelike geodesic, then it represents the world line of a free point mass. Let m be the mass of the point mass, then $U^a \equiv (\partial/\partial\tau)^a$ and $P^a \equiv mU^a$ are its 4-velocity and 4-momentum, respectively. Suppose p is a point on $\gamma(\tau)$, G is the static observer passing through p , Z^a is the 4-velocity of G at p (see Fig. 9.2), and $\xi^a = (\partial/\partial t)^a$ is the static Killing vector field, then it follows from $Z^a Z_a = -1$ that

$$Z^a = \chi^{-1} \xi^a, \quad (9.1.7)$$

where $\chi \equiv (-\xi^b \xi_b)^{1/2}$. From (6.3.17) we can see that $-Z_a P^a$ is the energy value obtained from the local measurement made by observer G on the point mass, which was denoted by E ; to avoid confusion, now we will denote it as E_{local} . The E defined by (9.1.4) can be rewritten as

$$E = -\xi_a U^a = -\frac{1}{m} \xi_a P^a = -\frac{\chi}{m} Z_a P^a = \frac{\chi}{m} E_{\text{local}}, \quad (9.1.8)$$

and thus $E \neq E_{\text{local}}$. If the geodesic $\gamma(\tau)$ reaches infinity, then $E \rightarrow E_{\text{local}}/m$ when $r \rightarrow \infty$, and hence E can be interpreted as the energy per unit mass obtained from the local measurement made on the point mass by a static observer at infinity. Since E is a constant on $\gamma(\tau)$, E_{local} is not a constant on it, i.e., it is E that is conserved in the motion of a free point mass instead of E_{local} . Therefore, E can be interpreted physically as the total energy (including gravitational potential energy) per unit mass of a free point mass. In contrast, E_{local} is the energy obtained from the local measurement made by a static observer G , which does not include the gravitational potential energy, and is not a conserved quantity along a geodesic. This can be interpreted physically as follows: although a free point mass does not experience any force other than gravity, the gravitational force does work on it when it is moving. Hence, as an energy excluding the gravitational potential energy, E_{local} is not a constant. Similarly, if $\gamma(\tau)$ is a null geodesic, then E can be interpreted as the total energy of the photon times \hbar^{-1} .

[Optional Reading 9.1.1]

For a timelike geodesic $\gamma(\tau)$ that does not reach infinity (e.g., the Earth rotating around the Sun), E is certainly still a constant; however, one cannot find a direct connection between it and an observer at infinity anymore. Although some literature still refers to this as the energy measured by observers at infinity, in this text we prefer the following perspective: the clearest meaning of the word “measure” in “a quantity measured by an observer” is a local measurement, which requires that the observer and the point mass (world lines) to be intersecting. When the world line of the point mass does not reach infinity, to add a modifier like “measured by an observer at infinity” we should specify a plan of indirect measurement (e.g., by means of the light emitted to infinity). However, in many cases it is difficult to find a practical plan to make an indirect measurement; the E of a timelike geodesics that does not reach infinity may be an example of this. We prefer to refer to E as the **energy** of the point mass whose world line is this geodesic without any additional modifier [see Wald (1984)]. It has the dimension of energy, and can even be interpreted physically as the sum of E_{local} and the gravitational potential energy, and thus totally deserves the designation “energy”. However, this is not the energy measured by any observer, and a modifier like “measured by observer at infinity” seems to be not necessary at all.

[The End of Optional Reading 9.1.1]

Now we come to the physical interpretation of the constant L . Suppose p is an arbitrary point on a timelike geodesic $\gamma(\tau)$, and Z^a is the 4-velocity of a static observer at p . Normalizing the coordinate basis vectors at p yields an orthonormal tetrad of the tangent space V_p at p :

$$(e_0)^a \equiv (1 - 2M/r)^{-1/2}(\partial/\partial t)^a = Z^a, \quad (e_1)^a \equiv (1 - 2M/r)^{1/2}(\partial/\partial r)^a, \\ (e_2)^a \equiv r^{-1}(\partial/\partial\theta)^a, \quad (e_3)^a \equiv r^{-1}(\partial/\partial\varphi)^a,$$

whose dual frame is

$$(e^0)_a = (1 - 2M/r)^{1/2}(dt)_a, \quad (e^1)_a = (1 - 2M/r)^{-1/2}(dr)_a, \\ (e^2)_a = r(d\theta)_a, \quad (e^3)_a = r(d\varphi)_a.$$

Let W_p be the 3-dimensional subspace orthogonal to Z_a in V_p , then $\{(e_1)^a, (e_2)^a, (e_3)^a\}$ and $\{(e^1)_a, (e^2)_a, (e^3)_a\}$ in the equations above are an orthonormal triad and its dual triad of W_p , respectively. The following discussion in the 3-dimensional language is relative to a static reference frame. Suppose U^a is the 4-velocity of a free point mass $\gamma(\tau)$ at p , $u^a \in W_p$ is its 3-velocity, then its 3-momentum is $p^a \equiv \gamma mu^a$, where m is the mass of the point mass and $\gamma \equiv -U^a Z_a$. Following the definition of the angular momentum \vec{j} in Euclidean space, i.e., $\vec{j} := \vec{r} \times \vec{p}$, we can define the angular momentum for a free point mass represented by $\gamma(\tau)$ as $j^a := \epsilon^a_{bc} \gamma m r^b u^c$, where $r^b \equiv r(e_1)^b$. Now let us show that the $|L|$ defined by (9.1.5) is the magnitude of the angular momentum per unit mass of a free point mass represented by $\gamma(\tau)$. Note that r^b is in the radial direction, and thus the radial component of u^c does not contribute to r^b . Therefore,

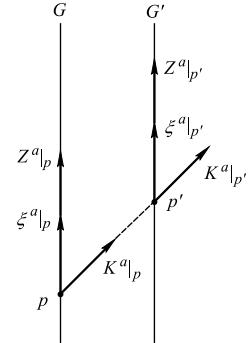
$$\begin{aligned}
j_a &= \gamma m \varepsilon_{abc} r^b u^3 (e_3)^c = \gamma m \varepsilon_{213} r u^3 (e^2)_a = -\gamma m r u^3 (e^2)_a, \\
|j| &= |\gamma m r u^3| = |\gamma m r u^a (e^3)_a| = |\gamma m r^2 u^a (\mathrm{d}\varphi)_a| = |m r^2 U^a (\mathrm{d}\varphi)_a| \\
&= |m r^2 (\partial/\partial\tau)^a (\mathrm{d}\varphi)_a| = m |r^2 \mathrm{d}\varphi/\mathrm{d}\tau| = m |L|,
\end{aligned} \tag{9.1.9}$$

where in the fourth equality we used the decomposition $U^a = \gamma(Z^a + u^a)$ and $Z^a(e^3)_a = 0$, and in the last step we used (9.1.5). Thus, (the absolute value of) L is the magnitude of the 3-momentum of a free point mass per unit mass relative to a static reference frame, or the angular momentum per unit mass for short. Similarly, if γ is a null geodesic, then L is the angular momentum of a photon times \hbar^{-1} .

9.2 Classical Experimental Tests of General Relativity

Basically, Einstein's original motivation for creating general relativity was purely theoretical. However, any physical theory must face the challenge of experimental verification after it comes out. We have seen in Sect. 7.9 that the direct detection of gravitational waves provided a strong confirmation of Einstein's theory one century after it came out. Nevertheless, during the formulation of general relativity, Einstein already made three important predictions early on by means of the vacuum Schwarzschild solution which could be compared with experiments (later on dubbed the three classical experimental tests). The earliest one (in 1907) is the gravitational redshift of light waves, and the other two are the precession of the perihelion of Mercury and the bending of starlight in the gravitational field of the Sun. The result of the perihelion precession calculation already agreed with the existing observational data, and the prediction of light deflection was also supported by observation very soon. However, due to the lack of experimental techniques for measuring extremely weak general relativity effects (including the gravitational redshift) with sufficient precision, the development in experimental researches of general relativity had been slow-going, or even almost stopped, for 45 years since the late 1910s. Since the 1960s, with the advancement of technology and the new discoveries of astronomical observations, the experimental verification of general relativity has entered its heyday; there appeared not only verifications of light deflection and gravitational redshift with a higher precision, but also a series of brand new experiments. It is safe to say that general relativity has passed all experimental tests so far, although many experiments with higher precision and difficulty are yet to be conducted. In this section, we will only discuss the three classical experimental tests proposed by Einstein. For the past, present, and future of the experimental tests of the relativistic theory of gravity, the reader may refer to Ni (2005; 2016).

Fig. 9.3 The derivation of the gravitational redshift in a stationary spacetime. G and G' are stationary observers



9.2.1 Gravitational Redshift

In this subsection we first discuss the gravitational redshift in a stationary spacetime, then, as an example, provide the explicit expression for the redshift of Schwarzschild spacetime. Under the geometric optics approximation, a light signal can be considered as propagating along a null geodesic (See the end of Sect. 7.2), and the angular frequency of a photon with a wave 4-vector K^a relative to an observer with a 4-velocity Z^a is [see (7.2.11)] $\omega = -K_a Z^a$.

Consider a stationary spacetime. Suppose G and G' are two observers in an arbitrary stationary reference frame. The photon emitted at p by G reaches G' at p' (see Fig. 9.3). Let Z^a be the 4-velocity of an observer, and K^a be the wave 4-vector of the photon, then the angular frequency of a photon at p and p' relative to the stationary observers at these points are, respectively,

$$\omega = -(K_a Z^a)|_p, \quad \omega' = -(K_a Z^a)|_{p'}. \quad (9.2.1)$$

The world lines of the stationary observers coincide with the integral curves of the Killing vector field ξ^a , and hence $\xi^a = \chi Z^a$, where χ can be obtained from $Z^a Z_a = -1$ to be $\chi \equiv (-\xi^b \xi_b)^{1/2}$. Thus, (9.2.1) becomes $\omega = [(-K_a \xi^a) \chi^{-1}]|_p$ and $\omega' = [(-K_a \xi^a) \chi^{-1}]|_{p'}$. Since the world lines of a photon is a geodesic whose tangent vector is K^a , and ξ^a is the Killing vector field, from Theorem 4.3.3 we can see that $K_a \xi^a$ is a constant on the curve, i.e., $(K_a \xi^a)|_p = (K_a \xi^a)|_{p'}$. Thus, it follows from (9.2.1) that¹

$$\frac{\omega'}{\omega} = \frac{\chi}{\chi'} \quad \text{or} \quad \frac{\lambda'}{\lambda} = \frac{\chi'}{\chi}, \quad (9.2.2)$$

where λ and λ' are the wave lengths corresponding to ω and ω' , respectively, and $\chi' \equiv (-\xi^b \xi_b)^{1/2}|_{p'}$. Now we will give the quantitative result for a static observer in

¹ There can be more than one null geodesic between two points p and p' in a stationary spacetime [See Sachs and Wu (1977) Exercise 7.3.2]. Equation (9.2.2) indicates that the redshift only depends on the points p and p' and has nothing to do with the null geodesics.

Schwarzschild spacetime as a specific example. Suppose $\{t, r, \theta, \varphi\}$ is the Schwarzschild coordinate system, then the timelike Killing vector field representing the staticity is $\xi^a = (\partial/\partial t)^a$, and hence

$$\chi^2 = -\xi^b \xi_b = -g_{ab} \left(\frac{\partial}{\partial t} \right)^a \left(\frac{\partial}{\partial t} \right)^b = -g_{00} = 1 - \frac{2M}{r}.$$

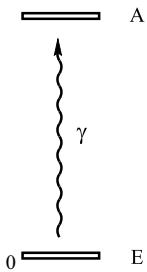
Plugging this into (9.2.2) yields

$$\lambda'/\lambda = (1 - 2M/r')^{1/2} (1 - 2M/r)^{-1/2}. \quad (9.2.3)$$

When $r' > r$ (i.e., light source is closer to the star than the receiver), we have $\lambda' > \lambda$, and thus the wave length of the light received by the receiver is longer than that when it was emitted, which is called a redshift. Since there is no relative motion between two stationary observers G and G' , the redshift can be interpreted as purely an effect of the gravitational field (curved spacetime). Hence, this effect is called a **gravitational redshift**, and $\chi \equiv (-\xi^b \xi_b)^{1/2}$ is called the (gravitational) **redshift factor**.

The magnitude of a redshift can be described by the relative redshift parameter (or simply **redshift**) $z \equiv (\lambda' - \lambda)/\lambda$. Calculation indicates that when the light emitted from the Sun arrives at the Earth (regard the Sun as a source of the gravitational field), the relative redshift is only about 2×10^{-6} . In order to enhance the redshift, one can measure the light coming from a white dwarf. A white dwarf is a celestial body which has a much higher density than a normal star (see Sect. 9.3.2 for details). Due to its high density, its surrounding gravitational field is way stronger than that of the Sun. The redshift of the light coming from a white dwarf can be dozens of times larger than the redshift of the light from the Sun. After general relativity was published, people have measured the redshift of the light from white dwarves a few times, but the results were not sufficiently precise to confirm the prediction of the theory. The first successful gravitational redshift experiment with high precision was done by R. V. Pound and G. A. Rebka Jr. using the Mössbauer effect in 1960. In 1960, R. L. Mössbauer discovered that some nuclei (e.g., ^{57}Fe) can emit γ -rays with very narrow (very sharp) linewidth under certain conditions, and crystals containing this kind of nucleus can have resonance absorption of γ -rays at this frequency with very high selectivity. Assuming that the frequency of this kind of γ -ray is changed slightly for some reason, the absorption by the crystal will be significantly reduced. This provides a powerful tool for measuring the extremely weak gravitational redshift caused by the Earth's gravitational field. Place two pieces of such a crystal at different heights on the surface of the Earth, the lower one (E in Fig. 9.4) as the emitter, and the higher one (see A in the figure) as the receiver. Although the redshift calculated based on the height difference between them (12.5 m) is merely 1.36×10^{-15} , the absorption rate of the γ -ray emitted by E to A still decreases due to the weak gravitational redshift of the γ -rays. To confirm and measure this decrease, one can let A move towards E at a constant speed, and use the “blueshift” (wave-

Fig. 9.4 Measuring the gravitational redshift near the ground using the Mössbauer effect



length decreases) due to the Doppler effect to offset the gravitational redshift. When the rate is adjusted to an appropriate value (only 3×10^{-7} m/s), the absorption rate will reach the maximum value. Then the value of the gravitational redshift can be measured. The precision of this experiment is very high (the relative uncertainty is about 1%), and the results obtained agree well with the theoretical values. Since then there were also experimental tests with higher precisions being done [the reader may refer to Will (2018)].

9.2.2 Perihelion Precession of Mercury

According to Newtonian mechanics, the orbit of a planet is an ellipse with the sun as a focus. However, the observational results are slightly divergent from this. Take Mercury, which is the closest to the Sun, as an example. Although in each period its orbit is very close to an ellipse, the major axes of two “ellipses” in two adjacent periods do not coincide, which is indicated by the slight change of its perihelion. As time goes on, due to the effect of accumulation, the slow rotation of the long axis of the “ellipse” (and thus the perihelion) around the sun becomes observable. This phenomenon is called the **precession** of the perihelion. Before the advent of general relativity, the precession rate of Mercury’s perihelion had already been measured as about 5600" per century (" stands for arcseconds). People have studied this in depth and discovered many possible causes (including the influence from the other planets). It was found that the precession rate caused by all these factors is 5557" per century, and there is still 43" per century that cannot be explained. This is the famous “43-second problem”. Based on general relativity, Einstein took Mercury as a free point mass in a curved spacetime caused by the Sun. His approximate calculation of a timelike geodesic in Schwarzschild spacetime naturally leads to the conclusion that the orbit of Mercury is not a closed curve, and the precession rate of its perihelion is exactly 43" per century. This result has greatly strengthened people’s confidence in general relativity. Now we will introduce the derivation of the perihelion precession in general relativity.

Suppose there are only the Sun and Mercury in the solar system and the gravitational field of Mercury can be neglected, i.e., we only discuss the motion of Mercury

under the action of the Sun's gravitational field (external gravitational field). First we discuss this using Newton's theory of gravity. Let the masses of the Sun and Mercury be M and m , respectively, then the gravitational potential energy of Mercury is

$$U(r) = -Mm/r \quad (\text{this text uses the system of geometrized units, where } G = 1). \quad (9.2.4)$$

Take the spherical coordinate system such that the orbit of Mercury is on the equatorial plane ($\theta = \pi/2$, which is always possible, see Sect. 9.1), then the velocity of Mercury has only a radial component $u_r = dr/dt$ and a tangential component $u_\varphi = r d\varphi/dt$, and hence the kinetic energy is $m(u_r^2 + u_\varphi^2)/2$. From the law of conservation of mechanical energy we have

$$\frac{1}{2}m(u_r^2 + u_\varphi^2) + U(r) = A, \quad (9.2.5)$$

where the constant A is the total mechanical energy of Mercury. Suppose the angular momentum of Mercury per unit mass is $|L|$, then

$$L = ru_\varphi = r^2 \frac{d\varphi}{dt}. \quad (9.2.6)$$

From (9.2.4), (9.2.5) and (9.2.6) we can find by calculation that

$$\left(\frac{dr}{d\varphi} \right)^2 + r^2 = \frac{2Mr^3}{L^2} + \frac{2Ar^4}{mL^2}. \quad (9.2.7)$$

Let $\mu \equiv r^{-1}$, then $\mu \neq 0$, and hence the above formula becomes

$$\left(\frac{d\mu}{d\varphi} \right)^2 + \mu^2 = \frac{2A}{mL^2} + \frac{2M}{L^2}\mu. \quad (9.2.8)$$

Taking the derivative with respect to φ yields $\frac{d\mu}{d\varphi} \left(\frac{d^2\mu}{d\varphi^2} + \mu - \frac{M}{L^2} \right) = 0$. Thus, either $\frac{d\mu}{d\varphi} = 0$ (round orbit), or

$$\frac{d^2\mu}{d\varphi^2} + \mu = \frac{M}{L^2}. \quad (9.2.9)$$

The solution to the above equation is

$$\mu(\varphi) = \frac{M}{L^2}[1 + e \cos(\varphi - \varphi_0)], \quad (9.2.10)$$

where e and φ_0 are constants of integration. Without loss of generality, take $\varphi_0 = 0$, then

$$\mu(\varphi) = \frac{M}{L^2}(1 + e \cos \varphi). \quad (9.2.11)$$

This is the equation of a conic section, with e as the eccentricity. Plugging (9.2.11) and its derivative back to (9.2.8) yields

$$e^2 = 1 + \frac{2AL^2}{mM^2}. \quad (9.2.12)$$

When $0 \leq e < 1$ this is an ellipse, and $d\mu/d\varphi = 0$ (round orbit) has been included in this as the special case of $e = 0$.

However, general relativity provides a slightly different result. Let $\kappa = 1$ (timelike geodesic). Dividing (9.1.6) by $(d\varphi/d\tau)^2$, and using (9.1.5) we can find by calculation that

$$\left(\frac{dr}{d\varphi}\right)^2 - \frac{E^2 r^4}{L^2} + r^2 \left(1 + \frac{r^2}{L^2}\right) \left(1 - \frac{2M}{r}\right) = 0. \quad (9.2.13)$$

Again let $\mu \equiv r^{-1}$, the above equation turns into

$$\left(\frac{d\mu}{d\varphi}\right)^2 + \mu^2 = \frac{1}{L^2} (E^2 - 1) + \frac{2M}{L^2} \mu + 2M\mu^3. \quad (9.2.14)$$

Taking the derivative with respect to φ yields

$$\frac{d^2\mu}{d\varphi^2} + \mu = \frac{M}{L^2} + 3M\mu^2. \quad (9.2.15)$$

Comparing this with (9.2.9) we find an additional term $3M\mu^2$ (general relativity correction term). Since the r of Mercury is way larger than the M of the Sun,² i.e., $M/r \ll 1$, the correction term $3M\mu^2 = (3M/r)\mu \ll \mu$, and thus one can manage to find an approximate solution. The solution (9.2.11) in Newton's theory of gravity can be viewed as the zeroth order approximation, denoted by $\mu_0(\varphi)$ for clarity, i.e.,

$$\mu_0(\varphi) = \frac{M}{L^2} (1 + e \cos \varphi). \quad (9.2.16)$$

Plugging this zeroth-order approximate solution into the second term on the right-hand side of (9.2.15), we obtain an equation that the first-order approximate solution $\mu_1(\varphi)$ should satisfy, i.e.,

$$\frac{d^2\mu_1}{d\varphi^2} + \mu_1 = \frac{M}{L^2} + 3M\mu_0^2 = \frac{M}{L^2} + \frac{3M^3}{L^4} (1 + 2e \cos \varphi + e^2 \cos^2 \varphi). \quad (9.2.17)$$

It is not difficult to verify that its solution is

² When doing the quantitative calculation, it is better to go back to the International System of Units (SI), i.e., to fill in the physical constants G and c . From Appendix A one can see that M/r is actually $(GM/c^2)/r$. The mass of the Sun M corresponds to $GM/c^2 \cong 1.5$ km, while the distance between the perihelion of Mercury and the Sun is about 5×10^7 km, and hence $(GM/c^2)/r \ll 1$.

$$\mu_1(\varphi) = \mu_0(\varphi) + \frac{3M^3}{L^4} \left[1 + e\varphi \sin \varphi + e^2 \left(\frac{1}{2} - \frac{1}{6} \cos 2\varphi \right) \right]. \quad (9.2.18)$$

What we care about is the perihelion. For $\mu_0(\varphi)$, the values of φ of the perihelion are $0, 2\pi, \dots$. Although there are many differences between the expressions of $\mu_1(\varphi)$ and $\mu_0(\varphi)$, the values of φ of the perihelion will not change if the term $e\varphi \sin \varphi$ is missing. Only this term can deviate Mercury from a closed orbit, which leads to the precession of the perihelion, and the precession angle increases as the value of φ increases (the effect will accumulate). Therefore, when we only care about the perihelion precession, we can neglect the other terms inside the square brackets in (9.2.18) except $e\varphi \sin \varphi$ and write it as [where $\mu_0(\varphi)$ has been substituted by (9.2.16)]

$$\mu_1(\varphi) = \frac{M}{L^2} \left[1 + e(\cos \varphi + \frac{3M^2}{L^2} \varphi \sin \varphi) \right]. \quad (9.2.19)$$

Since $M/L^2 \sim \mu$ [see (9.2.16)], we have $M^2/L^2 \sim M\mu = M/r \ll 1$. Let

$$\varepsilon \equiv \frac{3M^2}{L^2}, \quad (9.2.20)$$

then $\cos \epsilon\varphi \cong 1$, $\sin \epsilon\varphi \cong \epsilon\varphi$, and thus it follows from (9.2.19) that

$$\frac{1}{r(\varphi)} \cong \mu_1(\varphi) \cong \frac{M}{L^2} [1 + e \cos(\varphi - \varepsilon\varphi)]. \quad (9.2.21)$$

This indicates that the orbit of Mercury is approximately an ellipse. Although the right-hand side of (9.2.21) is still a periodic function, the period is not 2π as in (9.2.16). The perihelion is the point with the smallest r , i.e., the point where $\cos(\varphi - \varepsilon\varphi) = 1$. $\varphi = 0$ is certainly a perihelion; however, when $\varphi = 2\pi$,

$$\cos(\varphi - \varepsilon\varphi) = \cos(2\pi - 2\pi\varepsilon) \neq 1.$$

Suppose $\hat{\varphi}$ is the value of φ satisfying $\cos(\hat{\varphi} - \varepsilon\hat{\varphi}) = 1$ that is the closest to 2π , then it is not difficult to show that (neglecting the higher order term $2\pi\varepsilon^2$)

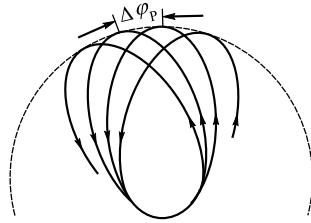
$$\hat{\varphi} \cong 2\pi + 2\pi\varepsilon. \quad (9.2.22)$$

Thus, the precession angle of the perihelion of Mercury in each period is (see Fig. 9.5)

$$\Delta\varphi_P \cong 2\pi\varepsilon = \frac{6\pi M^2}{L^2}. \quad (9.2.23)$$

The discussion above is valid for any planet. Plugging in the specific data one can obtain that the precession rate of the perihelion of Mercury is $43''$ per century.

Fig. 9.5 The precession angle $\Delta\varphi_p$ of the perihelion (aphelion) of Mercury in each period (with obvious exaggeration)



9.2.3 Light Deflection

When a light ray from a distant star that hits the ground after passing by the Sun, it will be bent due to the effect of the Sun's gravitational field. This is an important prediction of general relativity. In this section we introduce the derivation of this prediction. In the 4-dimensional language, the world line of a photon is a null geodesic. Let $\kappa = 0$ in (9.1.6), then using a method similar to the derivation of (9.2.13), it is not difficult to derive that

$$\left(\frac{dr}{d\varphi}\right)^2 - \frac{E^2 r^4}{L^2} + r^2 \left(1 - \frac{2M}{r}\right) = 0. \quad (9.2.24)$$

Again let $\mu \equiv r^{-1}$, then the equation above becomes

$$\left(\frac{d\mu}{d\varphi}\right)^2 + \mu^2 = \frac{E^2}{L^2} + 2M\mu^3. \quad (9.2.25)$$

Taking the derivative with respect to φ yields

$$\frac{d^2\mu}{d\varphi^2} + \mu = 3M\mu^2. \quad (9.2.26)$$

When $M = 0$ (flat spacetime), the general solution to (9.2.26) is

$$\mu(\varphi) = \frac{1}{l} \sin(\varphi + \alpha), \quad (9.2.27)$$

where l and α are constants of integration. Suppose the photon is at infinity when $\varphi = 0$, i.e., $\mu(0) = 1/r(0) = 0$, then $\alpha = 0$, and hence

$$\mu(\varphi) = \frac{1}{l} \sin \varphi. \quad (9.2.28)$$

This is a straight line equation in 2-dimensional Euclidean space expressed in a polar coordinate system $\{r, \varphi\}$. To see this, we take $r = 0$ as the origin of the Cartesian coordinate system $\{x, y\}$, then

$$x = r \cos \varphi , \quad (9.2.29)$$

$$y = r \sin \varphi = \frac{1}{\mu} \sin \varphi = l = \text{constant} , \quad (9.2.30)$$

where in the third equality we used (9.2.28). Thus, the spatial trajectory of a photon is a straight line whose distance from the origin is l (see Fig. 9.6). Note that both r and φ are changing along this straight line (y is a constant). Since the range of r is $(0, \infty)$, (9.2.29) indicates that the range of x is $(-\infty, \infty)$. To discuss the deflection of starlight, obviously we cannot take $M = 0$. However, since $M/r \ll 1$, finding the first-order approximate solution is sufficient for us. Taking the $\mu(\varphi)$ in (9.2.28) as the zeroth-order approximate solution $\mu_0(\varphi)$ and plugging it into the right-hand side of (9.2.26), we obtain the differential equation satisfied by $\mu_1(\varphi)$:

$$\frac{d^2 \mu_1}{d\varphi^2} + \mu_1(\varphi) = \frac{3M}{l^2} \sin^2 \varphi . \quad (9.2.31)$$

It is not difficult to verify that the solution to (9.2.31) is as follows:

$$\mu_1(\varphi) = \frac{1}{l} \sin \varphi + \frac{M}{l^2} (1 - \cos \varphi)^2 . \quad (9.2.32)$$

From the above equation we know that $\mu_1(0) = 0$, i.e., $r(0) = \infty$, which indicates that when the φ -coordinate of a photon is zero, it is infinitely far from the Sun (the r of a distant star can be regarded as ∞). However, (9.2.32) and (9.2.28) being different indicates that the photon is “heading” towards different directions when it is coming close to and leaving the Sun: it follows from (9.2.29) that $\mu_1(\pi) = 0$, which indicates that the φ -coordinate is π when the photon is going away from the “Sun” (this “Sun” has $M = 0$); however, it follows from (9.2.32) that $\mu_1(\pi) \neq 0$, and so we expect it to leave the Sun in a direction $\pi + \beta$ which is slightly different from π , i.e., $\mu_1(\pi + \beta) = 0$. To find the deflection angle β (see 9.7), by plugging $\varphi = \pi + \beta$ into (9.2.32) and using $\mu_1(\pi + \beta) = 0$ we obtain

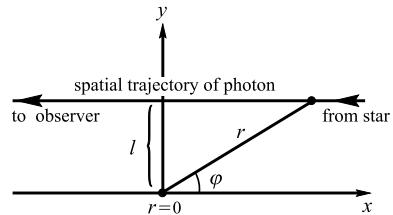
$$0 = \mu_1(\pi + \beta) = \frac{1}{l} \sin(\pi + \beta) + \frac{M}{l^2} [1 - \cos(\pi + \beta)]^2 .$$

The fact that β is small leads to $\sin(\pi + \beta) \cong -\beta$, $\cos(\pi + \beta) \cong -1$. Plugging them into the above equation yields

$$\beta \cong \frac{4M}{l} . \quad (9.2.33)$$

The above equation indicates that the deflection angle β increases as l decreases. The minimum value of l is equal to the radius of the Sun. Plugging this into (9.2.33) as the value of l [after adding the physical constants G and c , (9.2.33) becomes $\beta \cong 4GM/lc^2$], we find $\beta = 1.75''$. This is the quantitative prediction of general

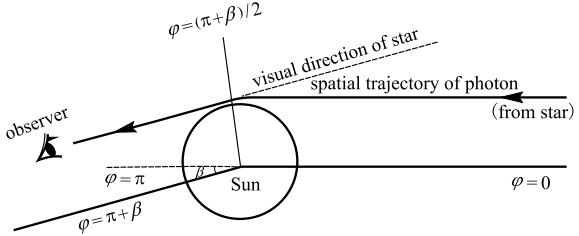
Fig. 9.6 The spatial trajectory of a photon in flat spacetime



relativity for the deflection angle of starlight. In order to verify this prediction by observation, we can try to photograph the apparent position of the star when the starlight is deflected by the Sun, and compare it with the actual position of the star photographed six months later (or ago) when the Earth turned to the other side of the Sun. However, it is not easy to observe the apparent position of a star, since the Sun is much closer to the Earth than the star we want to observe, and the starlight cannot be seen at all among the Sun's sunlight. (You cannot "watch the stars in the daytime"!) Then people came up with the idea of using a total solar eclipse. During a total solar eclipse, the sunlight is blocked by the Moon between the Sun and the Earth, but the light from a distant star can "bypass" the Sun and reach the Earth. Soon after World War I, two expedition teams set off from the United Kingdom to Brazil and Africa to observe the total solar eclipse on March 29th, 1919. The observation results of the two teams were, respectively, 1.13 ± 0.07 times and 0.92 ± 0.17 times of the theoretical prediction, which is considered as an important support of the theory. The announcement in many European and American newspapers attracted the attention of the war-weary public, and also made Einstein prestigious. However, Einstein responded quite calmly to this. He believed in his own theory so much (based on its elegance and internal self-consistency) that he once replied "Then I would feel sorry for the dear Lord." to the question what if the observation outcomes had not agreed with the theory. Note that Newton's theory of gravity can also predict the bending of starlight by the Sun, but the deflection angle is half of the predicted value of general relativity. The results of the 1919 UK teams indeed favored Einstein over Newton, but they were not of high precision. Although there were a few more observations for total solar eclipses in the next several decades, which continued to give mild support to general relativity, the improvement in accuracy was very little due to various reasons, especially the weather. In modern times, with technological advances, people can test the bending of distant quasar light by Jupiter and also the bending of radio waves, and the measurements are now quite a bit more precise.³ Therefore, we do have now pretty high precision results for light deflection which provide a strong support for general relativity [for details see Will (2018)].

³ For example, an analysis based on the very-long-baseline interferometry (VLBI) database gives a result which is 0.99983 ± 0.00045 times the predicted value of general relativity, where the standard error is reduced to 4.5×10^{-4} , see Shapiro et al. (2004).

Fig. 9.7 The Sun warps the spacetime, and the spatial trajectory of the starlight is deflected, where the deflection angle is obviously exaggerated



9.3 Spherical Stars and Their Evolution

9.3.1 Interior Solutions for Static Spherical Stars

In this subsection we discuss the interior spacetime metric and internal states of a static spherically symmetric star. The matter field inside a star can be regarded rather precisely as a perfect fluid, whose energy-momentum tensor is

$$T_{ab} = (\rho + p)U_a U_b + p g_{ab}. \quad (9.3.1)$$

The star being static means that every comoving observer inside the star can be regarded as a static observer, whose 4-velocity U^a is parallel to the static Killing vector field $\xi^a = (\partial/\partial t)^a$. Again, choose the Schwarzschild coordinate system, the line element can then still be expressed by (8.3.2). From $U^a U_a = -1$ and $\xi^a \xi_a = g_{00} = -e^{2A}$, we get

$$U^a = e^{-A}(\partial/\partial t)^a \quad \text{and} \quad U_a = -e^A(dt)_a, \quad (9.3.2)$$

and hence it follows from (9.3.1) that the nonvanishing components of T_{ab} are as follows:

$$\begin{aligned} T_{00} &= T_{ab}(\partial/\partial t)^a(\partial/\partial t)^b = \rho e^{2A}, & T_{11} &= T_{ab}(\partial/\partial r)^a(\partial/\partial r)^b = p e^{2B}, \\ T_{22} &= T_{ab}(\partial/\partial \theta)^a(\partial/\partial \theta)^b = p r^2, & T_{33} &= T_{ab}(\partial/\partial \varphi)^a(\partial/\partial \varphi)^b = p r^2 \sin^2 \theta. \end{aligned}$$

The Einstein equations $R_{\mu\nu} - Rg_{\mu\nu}/2 = 8\pi T_{\mu\nu}$ can be rewritten as $R^\mu{}_\nu - R\delta^\mu{}_\nu/2 = 8\pi T^\mu{}_\nu$. Noticing that $g^{\mu\nu}$ has only diagonal components, we obtain

$$T^0{}_0 = g^{00}T_{00} = -\rho, \quad T^1{}_1 = g^{11}T_{11} = p, \quad T^2{}_2 = g^{22}T_{22} = p, \quad T^3{}_3 = g^{33}T_{33} = p. \quad (9.3.3)$$

On the other hand, from the nonzero $R_{\mu\nu}$ in (8.3.5)–(8.3.8), we can see that

$$R = 2e^{-2B}[-A'' + A'B' - A'^2 + 2r^{-1}(B' - A') - r^{-2}] + 2r^{-2},$$

and further find that the nonzero $R^\mu{}_\nu - R\delta^\mu{}_\nu/2$ are as follows:

$$\begin{aligned} R_0^0 - \frac{1}{2}R\delta_0^0 &= -e^{-2B}(2B'r^{-1} - r^{-2}) - r^{-2}, \\ R_1^1 - \frac{1}{2}R\delta_1^1 &= e^{-2B}(2A'r^{-1} + r^{-2}) - r^{-2}, \\ R_2^2 - \frac{1}{2}R\delta_2^2 &= e^{-2B}[A'' - A'B' + A'^2 + (A' - B')r^{-1}], \\ R_3^3 - \frac{1}{2}R\delta_3^3 &= e^{-2B}[A'' - A'B' + A'^2 + (A' - B')r^{-1}]. \end{aligned}$$

Plugging these together with (9.3.3) into $R^\mu_\nu - R\delta^\mu_\nu/2 = 8\pi T^\mu_\nu$, we see that there are only 3 independent equations as follows:

$$-8\pi\rho = -e^{-2B}(2B'r^{-1} - r^{-2}) - r^{-2}, \quad (9.3.4)$$

$$8\pi p = e^{-2B}(2A'r^{-1} + r^{-2}) - r^{-2}, \quad (9.3.5)$$

$$8\pi p = e^{-2B}[A'' - A'B' + A'^2 + (A' - B')r^{-1}]. \quad (9.3.6)$$

Equation (9.3.4) can be rewritten as

$$8\pi\rho r^2 = 2re^{-2B}B' - e^{-2B} + 1 = 1 - \frac{d}{dr}(re^{-2B}),$$

and hence by integration we get

$$re^{-2B(r)} = r - 2m(r) + C, \quad (9.3.7)$$

where C is the constant of integration, and the function $m(r)$ is defined as

$$m(r) := 4\pi \int_0^r \rho(x)x^2 dx. \quad (9.3.8)$$

If $C \neq 0$, then from (9.3.7) and (9.3.8) we can see that $e^{-2B} \rightarrow \infty$ when $r \rightarrow 0$; however, $e^{-2B} = g^{11}$, while it is unreasonable to have $g^{11} = \infty$ in the center ($r = 0$) of the star, and hence $C = 0$. Thus, it follows from (9.3.7) that

$$g_{11}(r) = e^{2B(r)} = \left[1 - \frac{2m(r)}{r}\right]^{-1}. \quad (9.3.9)$$

Suppose the radius of the star is R , then when $r > R$ the metric should be the vacuum Schwarzschild solution [see (8.3.18)]. The interior metric and exterior metric should be continuous at the surface ($r = R$) of the star. Plugging $r = R$ into (9.3.9) yields

$$g_{11}(R) = \left[1 - \frac{2m(R)}{R}\right]^{-1}.$$

On the other hand, it follows from the vacuum Schwarzschild solution that

$$g_{11}(R) = \left(1 - \frac{2M}{R}\right)^{-1}.$$

Comparing these two equations and using (9.3.8) we obtain

$$M = m(R) = 4\pi \int_0^R \rho(r)r^2 dr. \quad (9.3.10)$$

[Optional Reading 9.3.1]

It seems that (9.3.10) is the same as the relation between the mass M and density $\rho(r)$ of a star in Newtonian mechanics. However, things are not as simple as that. Since the space Σ_t at a time t of a static reference frame has a non-Euclidean geometry h_{ab} (similar to the discussion in the end of Sect. 8.3.2), whose 3-dimensional proper volume element (i.e., the volume element associated with h_{ab}) is

$$\epsilon = \sqrt{h}dr \wedge d\theta \wedge d\varphi = \left[1 - \frac{2m(r)}{r}\right]^{-1/2} r^2 \sin\theta dr \wedge d\theta \wedge d\varphi,$$

Therefore, when calculating the integral one cannot use $r^2 \sin\theta dr \wedge d\theta \wedge d\varphi$ as volume element as in the 3-dimensional Euclidean space. However, the M in (9.3.10) is the result of integrating $\rho(r)$ with $r^2 \sin\theta dr \wedge d\theta \wedge d\varphi$ as the volume element. From the mathematical perspective, the integral (9.3.10) in the 3-dimensional non-Euclidean space (Σ_t, h_{ab}) with $r^2 \sin\theta dr \wedge d\theta \wedge d\varphi$ as the volume element is somewhat strange, but this does not mean that the M in (9.3.10) is a weird quantity. In fact, as the only parameter of the Schwarzschild solution, the physical meaning of M is crystal clear: it is the total mass (total energy) of Schwarzschild spacetime, which includes the gravitational potential energy (see Chap. 12 for details). However, $\rho(r)$ is the energy density obtained from the local measurement made by a static observer inside the star, which contains the static energy density of each particle (mainly the nucleus) and internal energy (heat, pressure, etc.) density, *except* the gravitational potential energy. This is similar to the discussion about the difference between E and E_{local} in Sect. 9.1: the result of a local measurement made by an observer does not contain the energy contribution from the gravitational field. Therefore, the M including gravitational potential energy is surely not equal to the integral $\int \rho(r)\epsilon$, since the latter does not include the contributions from the gravitational field. Note particularly that

$$\begin{aligned} \int \rho(r)\epsilon &= \int \rho(r) \left[1 - \frac{2m(r)}{r}\right]^{-1/2} r^2 \sin\theta dr \wedge d\theta \wedge d\varphi \\ &= 4\pi \int_0^R \rho(r) \left[1 - \frac{2m(r)}{r}\right]^{-1/2} r^2 dr \\ &\stackrel{!!!}{=} 4\pi \int_0^R \rho(r)r^2 dr = M. \end{aligned}$$

The fact that $\rho(r)$ does not contain the contribution from the gravitational field is closely related to another fact, namely the gravitational field energy is non-local. To put it in a simple way, the so called non-locality of the gravitational field energy means that the energy density of the gravitational field is *meaningless*: there does not exist such a quantity, which can be reasonably interpreted as the energy density of the gravitational field (NB: compare with the fact that the energy density of an electromagnetic field has a clear meaning and an explicit

expression), see Chap. 12 for details. However, the non-locality of the gravitational field energy does not indicate that the gravitational field itself has no energy. An important result people found after a long and tortuous path of study is: for an asymptotically flat spacetime (physically corresponding to an isolated gravitational system), one can always define the notion of total energy, which contains all the energy contributions including that of the gravitational field. Applying this definition to Schwarzschild spacetime with a parameter M , one finds that M is exactly the total energy of this spacetime (as an asymptotically flat spacetime).

[The End of Optional Reading 9.3.1]

Plugging (9.3.9) into (9.3.5) yields

$$\frac{dA}{dr} = \frac{m(r) + 4\pi pr^3}{r[r - 2m(r)]}. \quad (9.3.11)$$

Under the Newtonian approximation, ① the 3-dimensional space of a static reference frame can be approximately regarded as Euclidean space, with $1 \cong g_{11} = [1 - 2m(r)/r]^{-1}$, and hence $m(r) \ll r$; ② $p \ll \rho$ leads to $pr^3 \ll \rho r^3 \sim m(r)$, and thus (9.3.11) can be approximated as

$$\frac{dA}{dr} \cong \frac{m(r)}{r^2}. \quad (9.3.12)$$

Since the Newtonian gravitational potential ϕ with spherical symmetry satisfies

$$\frac{d\phi}{dr} = \frac{m(r)}{r^2}, \quad (9.3.13)$$

we can see that A is, in a sense, the quantity corresponding to the Newtonian gravitational potential in a static spherically symmetric curved spacetime. Equation (9.3.13) is actually a manifestation of the Poisson equation $\nabla^2\phi = 4\pi\rho$ in Newton's theory of gravity in the spherically symmetric case. When we have spherical symmetry, $\nabla^2\phi = 4\pi\rho$ becomes

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d\phi}{dr} \right) = 4\pi\rho.$$

Integrating this we obtain

$$r^2 \frac{d\phi}{dr} = 4\pi \int_0^r \rho(x)x^2 dx = m(r),$$

which is exactly (9.3.13).

Until now, among the 3 undetermined equations, the only equation we have not dealt with is (9.3.6). By plugging (9.3.9) and (9.3.11) into (9.3.6), this equation is in principle solvable, but the calculation will be quite complicated. By dint of the following fact (for proof see Optional Reading 9.3.2) the calculation can be simplified: under the premise that (9.3.4) and (9.3.5) are satisfied, (9.3.6) is equivalent

to

$$(\partial/\partial r)^b \nabla^a T_{ab} = 0. \quad (9.3.6')$$

It follows from (9.3.1) that

$$\nabla^a T_{ab} = U_a U_b \nabla^a (\rho + p) + (\rho + p) (U^a \nabla_a U_b + U_b \nabla_a U^a) + \nabla_b p.$$

Noticing that $(\partial/\partial r)^b$ is orthogonal to U^a , we can see that (9.3.6') is equivalent to

$$0 = (\partial/\partial r)^b \nabla^a T_{ab} = (\rho + p) (\partial/\partial r)^b U_a \nabla^a U_b + (\partial/\partial r)^b \nabla_b p, \quad (9.3.14)$$

Also,

$$\begin{aligned} (\partial/\partial r)^b \nabla_b p &= dp/dr, \\ (\partial/\partial r)^b U_a \nabla^a U_b &= -U_b U^a \nabla_a (\partial/\partial r)^b = e^A (dt)_b e^{-A} (\partial/\partial t)^a \nabla_a (\partial/\partial r)^b \\ &= (dt)_b \Gamma^{\sigma}_{10} (\partial/\partial x^{\sigma})^b = \Gamma^0_{10} = dA/dr, \end{aligned} \quad (9.3.15)$$

where (5.7.2) (the equivalent definition of Christoffel symbols) is used in the third equality, and (8.3.4) is used in the fifth equality. Plugging the above equation and (9.3.15) into (9.3.14) yields

$$\frac{dp}{dr} = -(p + \rho) \frac{dA}{dr}. \quad (9.3.16)$$

Then, using (9.3.11) we obtain

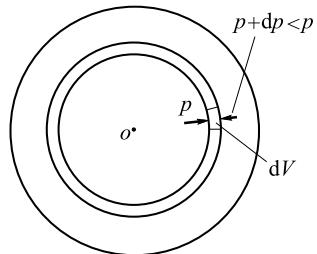
$$\frac{dp}{dr} = -(p + \rho) \frac{m(r) + 4\pi pr^3}{r[r - 2m(r)]}. \quad (9.3.17)$$

This is the famous **Oppenheimer-Volkoff (OV) equation of hydrostatic equilibrium**, whose Newtonian approximation is

$$\frac{dp}{dr} \underset{\text{Newtonian}}{\approx} -\frac{\rho m(r)}{r^2}, \quad (9.3.18)$$

where we used $p \ll \rho$, $m(r) \ll r$ and $pr^3 \ll \rho r^3 \sim m(r)$. The equation above is the well-known equation of hydrostatic equilibrium in Newtonian mechanics, which can be easily derived by means of Fig. 9.8. Due to the spherical symmetry, the gravitational force (self-gravity) from the star acting on any volume element in a thin spherical shell points to the center of the sphere. On the other hand, dV also experiences an outward force coming from the pressure gradient $dp/dr < 0$, and the star is in hydrostatic equilibrium when this force is equal to the self-gravity [i.e., satisfies (9.3.18)].

Fig. 9.8 The pressure gradient keeps the force on the volume element dV and the self-gravity in balance, from which (9.3.18) can be derived



In summary, the interior spacetime metric in a static spherically symmetric star is

$$ds^2 = -e^{2A(r)}dt^2 + \left[1 - \frac{2m(r)}{r}\right]^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2), \quad (9.3.19)$$

where the function $m(r)$ is defined by (9.3.8), and the function $A(r)$ needs to satisfy (9.3.11). A necessary and sufficient condition of hydrostatic equilibrium is (9.3.17).

The internal state inside a spherical star is determined by 4 functions $A(r)$, $m(r)$, $p(r)$ and $\rho(r)$, and there are only three equations they have to satisfy, namely (9.3.8), (9.3.11) and (9.3.17). In order to determine the internal state of a star, one also has to assign a fourth equation, called the equation of state. To put it in a simple way, an equation of state is a relation of the energy density ρ and pressure p represented by $f(\rho, p) = 0$ (where f is a certain specific function)⁴. After an equation of state is determined, there are only 3 undetermined functions $A(r)$, $m(r)$ and $p(r)$ remaining, which have to satisfy the differential equations (9.3.11), (9.3.17), and

$$\frac{dm(r)}{dr} = 4\pi\rho(r)r^2 \quad (9.3.20)$$

coming from (9.3.8). These 3 equations are all first-order differential equations, which can be solved exactly once the initial conditions $A(0)$, $m(0)$ and $p(0)$ are given. It follows from (9.3.8) that $m(0) \equiv 0$, and thus $m(0)$ does not need to be (and cannot be arbitrarily) assigned. After $A(0)$ (with adjustment later) and $p_0 \equiv p(0)$ are assigned, we can integrate the above-mentioned 3 differential equations from $r = 0$ to $p = 0$. [As long as the equation of state satisfies the following reasonable requirements: for all $p \geq 0$, we have $\rho \geq 0$, then the OV equation (9.3.17) assures automatically that the pressure decreases monotonically outwards.] The place where $p = 0$ is the surface of the star, whose corresponding value of r is the radius R of the

⁴ Generally speaking, the pressure p is not only a function of the density ρ , but also depends on the specific entropy (i.e., the average entropy per nucleus) and the chemical components of the star. Only when the specific entropy and chemical components are the same everywhere inside the star can p be solely a function of ρ , and the equation of state be expressed as $f(p, \rho) = 0$. The specific entropy of a normal star (including the Sun) is not everywhere the same. However, the specific entropy inside a white dwarf or neutron star, which will be discussed later, can be considered as vanishing everywhere. The discussion in the main text is valid for the study of these “abnormal celestial bodies”.

star, and $m(R)$ is the total mass (energy) M of the star (including the gravitational potential energy!). After having R we need to come back and modify the value of $A(0)$ (by adding a constant) in order to have it satisfy the condition at the surface of the star connecting the vacuum solution outside the sphere, i.e.,

$$e^{2A(R)} = 1 - \frac{2M}{R}. \quad (9.3.21)$$

Therefore, by assigning a value of p_0 one can determine a set of functions $A(r)$, $m(r)$ and $p(r)$, and the internal state and metric is then completely determined. For an equation of state in the real world, the exact solutions of equations like (9.3.17) are hard to be find, and thus a numerical method is used. However, for an idealized equation of state, we can perform the integral analytically. The simplest and most useful idealization is the following equation of state: $\rho = \text{constant}$. This is actually a very special equation of state whose energy density ρ is independent of the pressure. Although this is not a perfect model of a star, it can still be regarded as a first-order approximation of a small star whose pressure is not that high. Then (9.3.8) becomes

$$m(r) = \frac{4\pi\rho r^3}{3}. \quad (9.3.22)$$

The equation above holds for both general relativity and Newton's theory of gravity. For Newton's theory of gravity, (9.3.18) can be simplified as $\frac{dp}{dr} = -\frac{4\pi}{3}\rho^2r$ when ρ is a constant. After the initial value p_0 is assigned, the unique solution is $p(r) = -\frac{2}{3}\pi\rho^2r^2 + p_0$, and the radius R of the star can be determined by $p(R) = 0$:

$$0 = p(R) = -\frac{2}{3}\pi\rho^2R^2 + p_0.$$

Thus, p_0 can then be expressed in terms of R :

$$p_0 = \frac{2}{3}\pi\rho^2R^2, \quad (9.3.23)$$

and hence $p(r)$ can also be expressed in terms of R as

$$p(r) = \frac{2}{3}\pi\rho^2(R^2 - r^2). \quad (9.3.24)$$

When Newton's theory of gravity is not a good approximation, one needs to solve the OV equation (9.3.17). The solution found by Schwarzschild in 1916 is (and thus the metric inside a star with uniform density is called the **interior Schwarzschild solution**)

$$p(r) = \rho \frac{(1 - 2M/R)^{1/2} - (1 - 2Mr^2/R^3)^{1/2}}{(1 - 2Mr^2/R^3)^{1/2} - 3(1 - 2M/R)^{1/2}}, \quad (9.3.25)$$

and the corresponding central pressure is

$$p_0 = p(0) = \rho \frac{1 - (1 - 2M/R)^{1/2}}{3(1 - 2M/R)^{1/2} - 1}. \quad (9.3.26)$$

It is not difficult to show that (9.3.26) will approximately go back to (9.3.23) in Newton's theory of gravity when $R \gg M$ (Exercise 9.4).

Let $Y \equiv (1 - 2M/R)^{1/2}$, then (9.3.26) becomes

$$p_0 = \frac{\rho(1 - Y)}{3Y - 1}, \quad (9.3.27)$$

and it follows from $d\rho/dY < 0$ that p_0 increases as M/R increases. This is easy to understand since if M is larger, the self-gravity will be stronger, and so the pressure gradient for balancing the self-gravity will be greater, and the central pressure p_0 will be greater when R is fixed. In contrast, if M is fixed and R is smaller, then for the purpose of creating the pressure gradient we need, p_0 has to be higher as well. When M/R is large enough such that $Y = 1/3$, we have $p_0 \rightarrow \infty$, which indicates that equilibrium cannot be maintained no matter how large the central pressure is. Thus, the M/R of a static star with a uniform density has an upper limit, and from $Y = 1/3$ we can see that this upper limit is

$$(M/R)_{\max} = 4/9. \quad (9.3.28)$$

Of course, the M/R of a normal star is way smaller than this upper bound. To make a numerical evaluation, we should add the constant G/c^2 to M , i.e., substitute M by GM/c^2 . Take the Sun as an example, $GM_{\odot}/c^2 \cong 1.5$ km, $R_{\odot} \cong 7 \times 10^5$ km, and hence

$$\frac{GM_{\odot}/c^2}{R_{\odot}} \cong 2 \times 10^{-6} \ll \frac{4}{9}.$$

It follows from (9.3.22) that $M = 4\pi\rho R^3/3$, and eliminating R by using (9.3.28) yields

$$M_{\max} = \frac{4}{9} \frac{1}{\sqrt{3\pi\rho}}. \quad (9.3.29)$$

This is the maximum allowable mass of a star with uniform density ρ (note that there is no maximum allowable mass in Newton's theory of gravity). The existence of the upper mass limit in general relativity is not a result specifically for a star with uniform density. It can be proved that as long as one assumes $\rho(r) \geq 0$ and $d\rho/dr \leq 0$, the mass of any spherically symmetric static star with any radius R cannot exceed $4R/9$.

We mention in passing that, as we have emphasized in Sect. 8.10, when solving Einstein's equations, one should solve for the functions reflecting the matter field and the components of the metric simultaneously. In Example 1 of Sect. 8.10 we have pointed out that when the matter field is a perfect fluid, there are 16 undetermined functions $g_{\mu\nu}(x)$, $\rho(x)$, $p(x)$, $U^{\mu}(x)$ and 16 equations to be solved. The discussion in this section provides a specific example of that.

[Optional Reading 9.3.2]

Let $H_{ab} \equiv 8\pi T_{ab} - G_{ab}$, then H_{ab} only has diagonal components in the coordinate system $\{t, r, \theta, \varphi\}$. The fact that (9.3.4) and (9.3.5) holds is equivalent to $H_{00} = H_{11} = 0$. Now we will show that (9.3.6) is equivalent to (9.3.6') under the premise of (9.3.4) and (9.3.5). Noticing that $\nabla^a G_{ab} = 0$, we have

$$8\pi(\partial/\partial r)^b \nabla^a T_{ab} = (\partial/\partial r)^b \nabla^a H_{ab} = \nabla^a [(\partial/\partial r)^b H_{ab}] - H_{ab} \nabla^a (\partial/\partial r)^b. \quad (9.3.30)$$

Since $(\partial/\partial r)^b H_{ab} = H_{a1} = H_{11}(dr)_a = 0$, the first term on the right-hand side of the above equation vanishes. Let ∂_a be the ordinary derivative operator of the coordinate system $\{t, r, \theta, \varphi\}$, then the above equation becomes

$$\begin{aligned} 8\pi(\partial/\partial r)^b \nabla^a T_{ab} &= -H_b^a \left[\partial_a \left(\frac{\partial}{\partial r} \right)^b + \Gamma^b_{ac} \left(\frac{\partial}{\partial r} \right)^c \right] = -H^a_b \Gamma^b_{a1} \\ &= -(H^2_2 \Gamma^2_{21} + H^3_3 \Gamma^3_{31}) = -2H^2_2/r, \end{aligned}$$

where in the last step we used $H^2_2 = H^3_3$ and $\Gamma^2_{21} = \Gamma^3_{31} = 1/r$. Thus, $H^2_2 = 0$ [i.e., (9.3.6)] is equivalent to $(\partial/\partial r)^b \nabla^a T_{ab} = 0$.

[The End of Optional Reading 9.3.2]

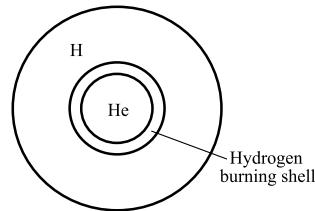
9.3.2 Stellar Evolution

In this subsection we introduce the formation and evolution of a spherically symmetric star. When the density is not very high, the gravitational field is not that strong, and Newton's theory of gravity is approximately applicable. General relativity is only necessary in the last part of this subsection. The predecessor of a star was a cloud of gas (mainly hydrogen) of inhomogeneous density. Where the density is higher, the gravity is stronger, which will attract more gas, and gradually forms a spherically symmetric gas cloud. The gravitational force (self-gravity) from the gas cloud acting on any volume element in any thin spherical shell (see Fig. 9.8) points to the center of the sphere, and so the entire gas cloud will contract under the action of self-gravity. This process transforms the gravitational potential into heat energy, and thus the temperature T keeps rising. According to the formula for the pressure of a classical ideal gas

$$p = k_B n T, \quad (k_B \text{ is the Boltzmann constant, } n \text{ is the number density}) \quad (9.3.31)$$

the pressure p in the gas cloud rises as T increases. Thus, the outward force on any thin spherical layer caused by the pressure gradient $d\rho/dr$ [see (9.3.18)] also increases with T , and it seems that the contraction may stop when the temperature is high enough. However, this is not possible without an energy source: since the temperature of the gas cloud is higher than its surroundings, it keeps radiating energy outwards. If the contraction stops, the temperature (and thus the pressure) will decrease, and the pressure difference between two sides of the thin shell cannot counterbalance the self-gravity. From the perspective of energy it also has to keep contracting, so that (a

Fig. 9.9 A rough sketch of the interior stellar core after the hydrogen in the core is burned up



part of) the gravitational potential energy keeps being converted into radiant energy. After the gas cloud contracts slowly for a period of time, the temperature and the density at the center is finally high enough to ignite a thermonuclear reaction. Near the center (a central sphere called the stellar core), the hydrogen is transformed into helium by thermonuclear fusion (which is the same reaction as in a hydrogen bomb explosion), and at the same time releases a huge amount of energy. This supplements the energy lost due to radiation (no need to rely on the gravitational potential energy conversion), and so the gas cloud will reach equilibrium and no longer contract. At this time, the gas cloud starts to become a star. The pressure gradient $d\rho/dr$ at any point in the gas cloud satisfies the stable equilibrium condition (9.3.18). The Sun is an example of an ordinary star. It has spent about 4.5 billion years in this stable state maintained by burning hydrogen into helium inside the stellar core, and can maintain this state for about another 5 billion years. One day, all the hydrogen in the stellar core will become helium, with only a thin layer of hydrogen around it still burning. The situation inside the star is roughly sketched in Fig. 9.9.

When the temperature of the stellar core has not reached the level of igniting helium nuclear fusion, the situation will be similar to the previous situation when it has not reached the point required to ignite hydrogen: the helium ball contracts again under the action of self-gravity and becomes hotter at the same time. This intensifies the burning of hydrogen in the surrounding thin layer, which leads to the expansion and cooling down of the outer part of the star, and turns it into a **red giant**. “Red” is due to the decrease of the surface temperature, while “giant” comes from its inflated size. The high temperature and density caused by the contraction of the helium sphere may reach the level of the nuclear fusion reaction that ignites helium (burning helium into carbon or oxygen), and the energy released will bring the stellar core to a stable equilibrium again. The duration of this balance maintained by helium combustion is much shorter than that of hydrogen combustion. When helium is burned into carbon (or oxygen), the stellar core will contract again. The fate of a star in its later years varies with its mass. For a star with a smaller mass (including the Sun), the contraction of the stellar core cannot provide enough temperature for carbon to undergo nuclear fusion, and thus it is no longer possible to maintain the equilibrium by nuclear energy. Is there any power strong enough to counterbalance the self-gravity? There does not exist such a power in classical physics. To prevent the contraction due to self-gravity, there must be a sufficiently large pressure gradient [which is represented by (9.3.18) in Newton’s theory of gravity, and (9.3.17) in general relativity].

A star is composed of hydrogen, helium and other elements. The high temperature in the star puts these atoms in ionized states. According to classical physics, this combination of ions and electrons can be regarded as an ideal gas. From (9.3.31) we can see that a high temperature is required in order to obtain a high pressure for a given density. Since the star keeps radiating energy, except for nuclear reactions, there does not exist such a mechanism that can provide energy for maintaining the high temperature. However, according to quantum physics, even a system at absolute zero temperature may have a considerable pressure. Take an electron gas for example. In classical physics, the average kinetic energy of the electrons is $3k_B T/2$; the average kinetic energy vanishes when $T = 0$, and so all electrons are in a state with zero energy. However, according to quantum physics, electrons are subject to the Pauli exclusion principle, i.e., any energy state can be occupied by at most two electrons (which have opposite spins and, hence, must be in different states). Therefore, when $T = 0$, the electrons on the one hand must “squeeze” into a state with the lowest possible energy; on the other hand, since each energy state can only be occupied by two electrons, electrons must fill up all the states with the energy values from zero all the way to a certain value E_F (only states with energy greater than E_F are all empty). E_F is called the **Fermi energy**, whose value increases as the density increases. This indicates that even at absolute zero, the electrons in the electron gas are not completely motionless as classical physics claims, they carry kinetic energy that is not due to thermal motion (but due to the exclusion principle). This kind of kinetic energy contributes to both pressure and energy density. An electron gas with $T = 0$ is called a (completely) **degenerate electron gas**, and the pressure caused by the above reasons is called the **electron degeneracy pressure**. At an ordinary density, the Fermi energy E_F is very small (for instance, the E_F of the electron gas in a common metal is only a few electronvolts), and the corresponding electron degeneracy pressure is negligible. However, the degeneracy pressure will have a considerable effect in the high density case. The high density caused by the second contraction of the stellar core when the hydrogen and helium are burnt up gives the electrons a rather high Fermi energy E_F . Although the temperature T in the stellar core is very high by the usual standard, due to the large E_F , we have $k_B T \ll E_F$, and thus the contribution of electrons to the pressure p due to the thermal motion is much smaller than that due to the kinetic energy of the electrons coming from the exclusion principle and high E_F . In this sense, it is not much different from the $T = 0$ case. So at this time, the electrons in the star can be regarded as a degenerate electron gas, whose degeneracy pressure may cancel the self-gravity, which will keep the star in equilibrium and never contract. This kind of stable star supported by the electronic degeneracy pressure is called a **white dwarf**. “Dwarf” means that it is much smaller than an ordinary star, and “white” is named due to the high temperature at its surface. Once an isolated star evolves into a white dwarf, there will be no important further evolution anymore. Since the temperature is higher than its surroundings, it will continuously radiate energy. Since there is no energy source, the radiation will cause the star’s temperature to decrease until it is equal to that of the surroundings, and so the star will no longer be visible (some literature refers to it as a “black dwarf”). The existence of white dwarfs has been confirmed by astronomical observations

dim and distant. Sirius B is the first white dwarf discovered by humans. Intuitively, the more massive a star is, the stronger the self-gravity it has; only a star with a sufficiently small mass can be supported by electronic degeneracy pressure and form a white dwarf. S. Chandrasekhar first found the upper mass limit of a white dwarf, $M_{\text{Ch}} \cong 1.3M_{\odot}$ [see Chandrasekhar (1939)]. This work along with his extraordinary contribution to astrophysics earned him the Nobel Prize in Physics in 1983. Optional Reading 9.3.3 will briefly introduce the derivation of the Chandrasekhar limit.

During its evolution, a star will eject matter which makes its mass decrease. We say that a white dwarf satisfies $M < M_{\text{Ch}}$, where M is the remaining mass. According to estimation, any star with its initial mass less than $6 \sim 8M_{\odot}$ will go through a red giant phase, eject a large amount of matter and become a white dwarf with its mass around $0.5 \sim 0.6M_{\odot}$.

If $M > M_{\text{Ch}}$, then the electron degeneracy pressure is not enough to maintain the equilibrium of the star, and the nuclear fusion reaction inside the stellar core will continue order by order until it is burned into iron and nickel. These are the most tightly bound nuclei (with the maximum average binding energy), so they do not release energy by nuclear fusion. Hence, the stellar core contracts sharply under the action of the self-gravity, and the density and temperature increase sharply. At this time the self-gravity is very strong, the Newtonian approximation (9.3.18) is no longer applicable, and (9.3.17) in general relativity must be used. For a given $\rho(r) > 0$, the right-hand side (absolute value) of (9.3.17) is always greater than that of (9.3.18); thus, to achieve an equilibrium in general relativity a greater central pressure is needed, and so the equilibrium is more difficult to achieve. At such a high temperature and high density, high-energy photons can break the iron-nickel nuclei into neutrons, protons, or light nuclei (photofission), and the electrons will also react with protons (electron capture) and form neutrons and neutrinos (the latter will run out of the star). Therefore, neutrons account for the vast majority in the stellar core. Neutrons are also fermions, which also obey the Pauli exclusion principle. When the nuclear density ($\sim 10^{17} \text{ kg}\cdot\text{m}^{-3}$) is reached, the Fermi energy E_F (divided by the Boltzmann constant k_B) of the neutrons is much higher than the temperature T in the star,⁵ and so it can be regarded as a degenerate neutron gas (i.e., $T \cong 0$), whose degeneracy pressure may also counterbalance the self-gravity, making the star reach a stable equilibrium. This kind of stable star supported by the neutron degeneracy pressure is called a **neutron star**. Since the density inside a neutron star reaches or even exceeds nuclear density, people's understanding of the equation of state under this kind of condition is far less accurate than that at lower densities, which makes it rather difficult to calculate the maximum mass of a neutron star. Different literature gives different values of this, and one can only roughly say that the upper mass limit of a neutron star is $2M_{\odot}$ (or $2 \sim 3M_{\odot}$). Since it reaches nuclear density, one may consider a neutron star as a “super-large atomic nucleus”. A neutron star is much smaller than a white dwarf. The typical radius of a neutron star is only on the order of 10 km, whereas a white dwarf has a radius between about 3,000 and 20,000

⁵ A more precise statement is: since it releases a large amount of high-energy neutrinos, a few seconds after the formation of the neutron star it has $E_F \gg k_B T$.

kilometers. A neutron star is a very special (and complex) celestial body, which has various “extreme” (abnormal) behaviors: a density up to nuclear density, unusually strong magnetic field (up to 10^{12} Gauss), very high-speed rotation (with frequency from 1 Hz to nearly 1000 Hz), high speed of sound which is close to the speed of light, superfluid in the interior.... Until today, it is still difficult to understand it thoroughly.

The first theoretical model of a neutron star was published by J. R. Oppenheimer and G. M. Volkoff in 1939. Since their article did not provide any observable physical effect, the study of neutron stars had been slighted for 28 years. The existence of neutron stars has been confirmed since the discovery of a pulsar in 1967. A pulsar is a signal source of periodic electromagnetic pulse signals measured on Earth, with a period about 1 s or less. The only persuasive explanation is: this is a rotating neutron star whose strong magnetic field on the surface leads to magnetic dipole radiation, and the combination of orientation of the radiation and the rotation of the neutron star lets the Earth receive an electromagnetic pulse signal (the electromagnetic pulse of the pulsar discovered in 1967 is a radio pulse). Only neutron stars (with small radius and strong surface gravity) can rotate at such a high angular velocity without “falling apart”.

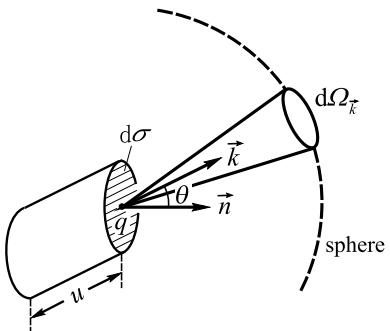
The stellar core contracts very sharply before it forms a neutron star, and thus this process is called **gravitational collapse**. Once the rapidly collapsing stellar core reaches a sufficient density and is stopped by the neutron degeneracy pressure, its high energy will appear as an outward shock wave, and bust out the outer material, forming a **supernova explosion** with a great energy. Pulsars have been found in two famous supernova remnants, the Crab Nebula and the Vela supernova remnant, which provides an important support for the above-mentioned theory. Ancient Chinese documents have extremely rich records of supernova explosions. For example, Volume Nine of *Zhi (Records)* in *Song Shi (History of Song)* published in 1346 recorded the supernova (SN1054) observed in 1054 AD (during the Northern Song Dynasty), which was particularly valued by modern international peers [the photo of one page of it can be found at the title page of Misner et al. (1973)]. The Crab Nebula is exactly the remnant of SN1054. The most recently observed supernova explosion visible to the naked eye on Earth was in 1987 (SN1987a). This supernova is located in a neighboring galaxy of the Milky Way—the Large Magellanic Cloud, which is about 160,000 light-years away from the Earth. The detailed mechanism of the supernova explosion is still a subject being studied in depth.

If the mass of a spherically symmetric star is still greater than the upper mass limit of a neutron star ($\sim 2M_{\odot}$) after ejecting matter, there will be no power to prevent its gravitational collapse. Then, it will contract without any restriction into a “singularity” with infinite density and curvature, and form a Schwarzschild black hole (see Sect. 9.4).

[Optional Reading 9.3.3]

This optional reading introduces the derivation of the formula of electron degeneracy pressure and the upper mass limit of a white dwarf. First we discuss the electron degeneracy pressure. Suppose x, y, z are the spatial coordinates of an electron, and k_x, k_y, k_z are the three coordinate components of the electron’s momentum, then $\{x, y, z; k_x, k_y, k_z\}$ is a coordinate system of the 6-dimensional phase space. A phase space can be divided into many quantum phase cells $dx dy dz dk_x dk_y dk_z$ (each phase cell corresponds to an energy level), and the

Fig. 9.10 Electrons inside an oblique cylinder that go through the area element $d\sigma$ in a unit time, from which one can calculate the pressure at $d\sigma$



volume of a phase cell is h^3 (where h is the Planck constant). Hence,

$$dxdydzdk_xdk_ydk_z = h^3. \quad (9.3.32)$$

Let $k \equiv (k_x^2 + k_y^2 + k_z^2)^{1/2}$, then the points whose values of k are in the range of $(k, k + dk)$ in the momentum space constitute a spherical shell with volume $4\pi k^2 dk$. Then, the points in the phase space representing states whose position is in $dxdydz$ and the value of k is in $(k, k + dk)$ constitute a shell with volume $4\pi k^2 dk dx dy dz$. Since the volume of each quantum phase cell is h^3 , there will be $4\pi k^2 dk dx dy dz / h^3$ phase cells in the shell. Since each cell corresponds to an energy level, and each energy level is occupied by at most two electrons, the number of electrons in a shell will not exceed $8\pi k^2 dk dx dy dz / h^3$. For a completely degenerate electron gas with $T = 0$, each energy level with $E \leq E_F$ has two electrons, and all the energy levels with $E > E_F$ are empty. Therefore, the number of electrons with their values of k in $(k, k + dk)$ per unit volume, denoted by $f(k)dk$, satisfies

$$f(k)dk = \begin{cases} 8\pi k^2 dk / h^3, & k < k_F \\ 0, & k > k_F \end{cases}, \quad (9.3.33)$$

where k_F is the **Fermi momentum** corresponding to E_F . Hence, the number density of electrons (the number of electrons per unit volume regardless of their momenta) is

$$n_e = \int_0^{k_F} \frac{8\pi k^2 dk}{h^3} = \frac{8\pi}{3h^3} k_F^3. \quad (9.3.34)$$

The predominant contribution to the mass density ρ in a star comes from the nuclei. Suppose the number density and mass of nuclei are n_N and m_N , then $\rho = n_N m_N$. Let $\mu \equiv n_N/n_e$ (for a star with hydrogen burnt up, $\mu \cong 2$), then

$$\rho = \mu n_e m_N, \quad (9.3.35)$$

where n_e is given by (9.3.34). To obtain the equation of state, one should also compute the degeneracy pressure p_{de} . Pressure is the stress per unit area, i.e., the force that the matter on the left side of an area element exerts on the matter on the right side, or the momentum exchanged through the area element per unit time (as the definition of force is the rate of change of momentum $d\vec{k}/dt$). This exchange of momentum is caused by the electrons going across the area from left to right or the other way around (each electron carries some certain momentum). Therefore, the pressure equals the vector sum of the momenta of the electrons going through per unit area per unit time. Suppose $d\sigma$ is an area element in the internal space of the star, whose normal vector is \vec{n} (see Fig. 9.10). First, consider an electron

with momentum \vec{k} whose corresponding velocity is \vec{u} , i.e., $\vec{k} = (1 - u^2)^{-1/2} m_e \vec{u}$ (where $u^2 \equiv \vec{u} \cdot \vec{u}$). Take an oblique cylinder with $d\sigma$ as the base and u as the generatrix length, whose generatrix is parallel to \vec{k} . Suppose θ is the angle between \vec{n} and \vec{k} , then the volume of the cylinder equals $u \cos \theta d\sigma$, and it follows from (9.3.33) that the number of electrons inside the cylinder whose values of k are in $(k, k + dk)$ is $f(k)u \cos \theta d\sigma dk$. So far we have not considered the direction of \vec{k} yet. Take a sphere with a point q in the area element $d\sigma$ as the center. Divide the right hemisphere into many area elements, each of which corresponds to a solid angle element, where the one with \vec{k} as the axis is denoted by $d\Omega_{\vec{k}}$. Since the solid angle corresponding to the whole sphere is 4π , the number of electrons inside the cylinder whose momenta are oriented within $d\Omega_{\vec{k}}$ and magnitudes are in $(k, k + dk)$ is only $f(k)u \cos \theta d\sigma dk d\Omega_{\vec{k}} / 4\pi$. These electrons (carrying momentum) are going through $d\sigma$ in every unit of time, and thus the normal component of the total momentum of the electrons going through $d\sigma$ per unit time satisfying the above-mentioned conditions [① the magnitudes are in $(k, k + dk)$; ② the directions are in $d\Omega_{\vec{k}}$] is

$$f(k)u \cos \theta d\sigma dk \frac{d\Omega_{\vec{k}}}{4\pi} k \cos \theta = f(k)u k \cos^2 \theta d\sigma dk \frac{d\Omega_{\vec{k}}}{4\pi}.$$

Hence, the total momentum of all electrons (regardless of the magnitudes and directions of \vec{k}) going through per unit area per unit time, i.e., the degeneracy pressure at $d\sigma$ is

$$p_{de} = \frac{1}{4\pi} \int_{\text{sphere}} \cos^2 \theta d\Omega_{\vec{k}} \int_0^\infty f(k)u(k)k dk = \frac{8\pi}{3h^3} \int_0^{k_F} k^3 u(k) dk, \quad (9.3.36)$$

where (9.3.33) is used in the second equality. Using $\vec{k} = (1 - u^2)^{-1/2} m_e \vec{u}$, one can rewrite (9.3.36) as

$$p_{de} = \frac{8\pi}{3h^3} \int_0^{k_F} \frac{k^4 dk}{(k^2 + m_e^2)^{1/2}}. \quad (9.3.37)$$

Then, using (9.3.34) and one can rewrite (9.3.35) as

$$k_F = h \left(\frac{3\rho}{8\pi\mu m_N} \right)^{1/3}. \quad (9.3.38)$$

Plugging this into (9.3.37) yields the explicit expression for the equation of state. This equation is quite complex, but some useful conclusions can be obtained by analyzing two extreme cases. When $m_e \gg k_F, u_F \ll 1$, the motion of electrons can be described by Newtonian mechanics, which is called the non-relativistic case; when $m_e \ll k_F, u_F \cong 1$, the motion of electrons must be characterized by special relativity, which is called the ultra-relativistic case. The non-relativistic condition $m_e \gg k_F$ and the ultra-relativistic condition $m_e \ll k_F$ can also be expressed as $\rho \ll \rho_C$ and $\rho \gg \rho_C$, respectively, where the **critical density** ρ_C is defined by $m_e = k_F$, which can be found explicitly from (9.3.38) as

$$\rho_C = \frac{8\pi\mu m_N m_e^3}{3h^3}. \quad (9.3.39)$$

Rewriting this in SI (adding c^3) and plugging in the specific values (take $\mu = 2$), we obtain

$$\rho_C = \frac{8\pi\mu m_N m_e^3 c^3}{3h^3} \cong 2 \times 10^9 \text{ kg} \cdot \text{m}^{-3},$$

and thus the critical density ρ_C is about 2×10^6 times the density of water. For $\rho \ll \rho_C$ (non-relativistic case), (9.3.37) gives approximately

$$p_{de} = \frac{8\pi k_F^5}{15h^3 m_e}. \quad (9.3.40)$$

Plugging in (9.3.38) along with the specific values in SI yields

$$p_{de} = \frac{1}{20} \left(\frac{3}{\pi} \right)^{2/3} \frac{h^2}{m_e m_N^{5/3}} \left(\frac{\rho}{\mu} \right)^{5/3} = 10^7 \left(\frac{\rho}{\mu} \right)^{5/3} \quad (\text{SI}). \quad (9.3.41)$$

Under the ultra-relativistic condition, (9.3.37) gives approximately

$$p_{de} = \frac{2\pi k_F^4}{3h^3}. \quad (9.3.42)$$

Plugging (9.3.38) in the above equation, rewriting it in SI (adding c) and plugging in the specific values, we obtain

$$p_{de} = \left(\frac{3}{\pi} \right)^{1/3} \frac{hc}{8m_N^{4/3}} \left(\frac{\rho}{\mu} \right)^{4/3} = 1.24 \times 10^{10} \left(\frac{\rho}{\mu} \right)^{4/3} \quad (\text{SI}). \quad (9.3.43)$$

Eqs. (9.3.41) and (9.3.43) can be unified as

$$p_{de} = K\rho^\gamma, \quad K = \text{constant}, \quad \gamma = \text{constant}. \quad (9.3.44)$$

A celestial body whose equation of state is (9.3.44) is called a **polytrope**. Under both of the extreme cases, a star constituted of a degenerate electron gas is a polytrope (but in between the two cases it is not), where for the non-relativistic case we have $\gamma = 5/3$, and for the ultra-relativistic case we have $\gamma = 4/3$. Suppose a star is a polytrope, we can rewrite the hydrostatic equilibrium condition (9.3.18) using (9.3.8) as

$$\frac{d}{dr} \left[\frac{r^2}{\rho(r)} \frac{dp(r)}{dr} \right] = -4\pi\rho(r)r^2. \quad (9.3.45)$$

Starting from this equation, using some calculation techniques [see Weinberg (1972) pp. 308–310], one finds the dependence of the radius R and the mass M of the star on the central density ρ_0 :

$$R = a_\gamma \rho_0^{(\gamma-2)/2}, \quad (9.3.46)$$

$$M = b_\gamma \rho_0^{(3\gamma-4)/2}, \quad (9.3.47)$$

where the constants a_γ and b_γ are related to γ ; for $\gamma = 5/3$ and $\gamma = 4/3$ they are, respectively,

$$\begin{aligned} a_{5/3} &= 6.3 \times 10^8 \mu^{-5/6}, & b_{5/3} &= 1.7 \times 10^{26} \mu^{-5/2}, \\ a_{4/3} &= 5.3 \times 10^{10} \mu^{-2/3}, & b_{4/3} &= 11.6 \times 10^{30} \mu^{-2}. \end{aligned} \quad (9.3.48)$$

Based on this we can further discuss white dwarfs. When the mass M of a star is small enough, $\rho_0 \ll \rho_C$, then (9.3.41) is valid everywhere inside the star, and the electron gas inside the whole star forms a polytrope with $\gamma = 5/3$. When the central degeneracy pressure equals the central pressure for keeping equilibrium, the star will be in an equilibrium state. The relation between the radius R and the mass M in equilibrium can be seen from (9.3.46), (9.3.47) (with $\gamma = 5/3$) and (9.3.48) as

$$R \propto M^{-1/3}. \quad (9.3.49)$$

Thus, the radius of a white dwarf with $\gamma = 5/3$ decreases as the mass increases. This seems to contradict our life experience and the experience from the planets; later we will give a rough explanation of this. If (9.3.49) always holds, then the electron degeneracy pressure can support stars of any mass, since one can always plug a value of M into (9.3.49) and find a radius R of a star in equilibrium. However, when the mass M is sufficiently large, the central pressure will be so large that the (special) relativistic effect of electrons has to be considered. Then, the star can no longer be regarded as a polytrope with $\gamma = 5/3$, and (9.3.49) no longer holds. In fact, since ρ_0 increases as M increases, the electrons near the center will be the first to reach the ultra-relativistic level. A spherical core which can be regarded as a polytrope with $\gamma = 4/3$ will appear in the star, and then it will gradually expand to the entire body. From (9.3.47)⁶ we can see that M is independent of ρ_0 when $\gamma = 4/3$, which is quite different from the case where $\gamma = 5/3$. When the entire star can be regarded as a polytrope with $\gamma = 4/3$, it follows from (9.3.48) that the value of this M (denoted by M_{Ch}) which is independent of ρ_0 is

$$M_{\text{Ch}} = b_{4/3} = 5.8 \times (2 \times 10^{30}) \mu^{-2} \quad (\text{SI}).$$

Noticing that $M_{\odot} = 2 \times 10^{30}$ in SI, we have

$$M_{\text{Ch}} = \frac{5.8}{\mu^2} M_{\odot}. \quad (9.3.50)$$

Equation (9.3.47) is derived under the condition of hydrostatic equilibrium. If the mass is larger than M_{Ch} , the star cannot be in equilibrium. In fact, from (9.3.49) and (9.3.44) we can see that the conclusion above can be interpreted as follows: as a rough evaluation, we assume that the star has a uniform density, then it follows from (9.3.23) that the central pressure for keeping the equilibrium is

$$p_{\text{grav}} \propto M^2 R^{-4}, \quad (9.3.51)$$

where p_0 is now denoted as p_{grav} to emphasize that this is the central pressure for counter-balancing the self-gravity. It follows from (9.3.44) that the degeneracy pressure provided by the degenerate electron gas is $p_{\text{de}} \propto M^{\gamma} R^{-3\gamma}$, and hence

$$\frac{p_{\text{grav}}}{p_{\text{de}}} \propto M^{2-\gamma} R^{3\gamma-4} = \begin{cases} M^{1/3} R, & \text{for } \gamma = 5/3, \\ M^{2/3}, & \text{for } \gamma = 4/3. \end{cases} \quad (9.3.52a)$$

$$(9.3.52b)$$

Suppose the electron gas in the star is in the non-relativistic case ($\gamma = 5/3$) and $M < M_{\text{Ch}}$, then from (9.3.52a) we know that there exists an R such that $p_{\text{grav}}/p_{\text{de}} = 1$, and the star is in equilibrium when its radius equals this value of R . If M increases slightly, then $p_{\text{grav}}/p_{\text{de}} > 1$, i.e., the self-gravity is slightly larger than the degeneracy pressure, and the star will contract to a smaller radius to reach equilibrium again. (This can be considered as a specific interpretation for the conclusion “a white dwarf with a greater mass has a smaller radius”.) However, if M is so large that the entire star has $\gamma = 4/3$, then it follows from (9.3.52b) that $p_{\text{grav}}/p_{\text{de}}$ is independent of R . Under this extreme circumstance, only when M equals a suitable value M_{Ch} can the star be in equilibrium. If $M < M_{\text{Ch}}$, then $p_{\text{grav}}/p_{\text{de}} < 1$, i.e., the degeneracy pressure is greater than the self-gravity, R will increase until it quits the ultra-relativistic case. In contrast, if $M > M_{\text{Ch}}$, then $p_{\text{grav}}/p_{\text{de}} > 1$, and the star will contract, which makes γ closer to exactly $4/3$. Then $p_{\text{grav}}/p_{\text{de}}$ will not change with the decrease of R , and hence the star can only continue contracting and cannot reach equilibrium under the support of the electron degeneracy pressure. Thus, M_{Ch} is indeed the upper mass limit of a white dwarf (whose character is that the electron degeneracy pressure keeps the equilibrium).

⁶ Here we still have $p \ll \rho$ and $m(r) \ll r$, and hence the Newtonian equation (9.3.18) and (9.3.45)–(9.3.48) derived from it are still applicable.

Since the interior of a white dwarf is mostly helium, carbon or oxygen, one can take $\mu = 2$ in (9.3.50) and obtain $M_{\text{Ch}} = 1.45M_{\odot}$. The discussion above is just a simplified version. Some more precise discussions and calculations provide M_{Ch} slightly smaller than this value, such as $M_{\text{Ch}} = 1.3M_{\odot}$.

[The End of Optional Reading 9.3.3]

9.4 The Kruskal Extension and Schwarzschild Black Holes

The line element of the vacuum Schwarzschild metric in the Schwarzschild coordinate system is

$$ds^2 = - \left(1 - \frac{2M}{r}\right) dt^2 + \left(1 - \frac{2M}{r}\right)^{-1} dr^2 + r^2 (d\theta^2 + \sin^2 \theta d\varphi^2). \quad (9.4.1)$$

When $r = 2M$, $g_{11} = \infty$ (singular); when $r = 0$, both g_{00} and g_{11} are singular. These places where $g_{\mu\nu}$ is singular (or degenerate) are called **singularities**. Note that the word “singularity” may be used to refer to both the property of being singular and the place that has the singularity.⁷ There are two reasons accounting for the appearance of a singularity: ① the metric tensor g_{ab} is well-behaved at this place, just some components are not well-behaved in certain coordinate systems; this is called a **coordinate singularity**, which can be removed by choosing a suitable coordinate system; ② The metric tensor g_{ab} itself is ill-behaved (singular) at this place; this is called a **true singularity** or **spacetime singularity**, which is really a thorny problem in general relativity. Later we will see that the singularity at $r = 2M$ is only a coordinate singularity, and a spacetime singularity exists only at $r = 0$. Denoting $r_S \equiv 2M$ (called the **Schwarzschild radius**, and adding the constants G and c , we get

$$r_S = \frac{2GM}{c^2} \cong \frac{3M}{M_{\odot}} \quad (\text{km}).$$

For the Sun, $r_S \cong 3$ km, which is far less than its radius. Since the external Schwarzschild solution does not apply to the interior of the Sun, there is no singularity problem for it (or any normal celestial bodies). However, for a spherically symmetric star which experiences gravitational collapse and turns into a black hole (Birkhoff’s theorem assures that the external spacetime geometry is described by the Schwarzschild metric), the singularity problem is of great significance.

⁷ Also note that a singularity may not be a point, since in the 4-dimensional language $r = 0$ (or $r = 2M$) represents a hypersurface instead of a point.

9.4.1 The Definition of a Spacetime Singularity

The concept of singularity is closely related to the divergence of physical quantities, which has existed long before general relativity came out. However, the problem of spacetime singularities in general relativity is much more troublesome than the singularity problem in any other physical theories (even the definition is more troublesome). The key point is that, in any theory that is not of general covariance, a background spacetime is given beforehand (e.g., Minkowski spacetime). Wherever the physical field we care about is divergent (or undefined), we say that this is a singularity of the physical field. For instance, in the 3-dimensional language, the electrostatic field strength $\vec{E} = Q\vec{r}/4\pi r^3$ of a point charge is undefined at $r = 0$ (i.e., $E \rightarrow \infty$ when $r \rightarrow 0$), and thus we say that this point is the singularity of the electrostatic field \vec{E} , or say that \vec{E} is singular at $r = 0$. However, things are different in general relativity. Since what we care about is the spacetime singularity, i.e., the singularity of the metric, the metric field plays a double role of both the background field and the physical field in this problem (it acts as both the stage and the actor). Following the definition of the singularity in an electrostatic field, it seems that one can define a spacetime singularity as follows: “a spacetime (M, g_{ab}) is said to be singular if $\exists p \in M$ such that g_{ab} is undefined (or divergent) at p , and the point p is called a spacetime singularity.” However, the spacetime itself, by definition, is a 4-dimensional manifold M equipped with a Lorentzian metric, and at each point of M the metric g_{ab} should be not only well-defined, but also well-behaved, such as being continuous and differentiable to some certain order. If there exists a point p in M such that $g_{ab}|_p$ is undefined, then p does not even belong to the spacetime (not a valid spacetime point), and thus the actual spacetime is (M', g_{ab}) , where M' is the result of eliminating the point p , i.e., $M' = M - \{p\}$. For example, if we use (M, g_{ab}) to represent Schwarzschild spacetime, then it should be make clear that all the points with $r = 0$ do not belong to M . It seems that the definition of spacetime singularity can be modified as: “a spacetime is said to be singular if some of its region is eliminated.” But the problem is how to determine whether “one or some regions are eliminated”. Here we introduce an ingenious method. Take Minkowski spacetime as an example. Suppose $\gamma(\lambda)$ is an arbitrary inextendible geodesic in $(\mathbb{R}^4, \eta_{ab})$ (meaning both ends have been extended until they cannot be extended anymore), then its affine parameter λ takes values from $-\infty$ to $+\infty$ [i.e., $\gamma(\lambda) \in \mathbb{R}^4, \forall \lambda \in (-\infty, \infty)$]. If we eliminate a point p of $\gamma(\lambda)$, there will be a “hole” left in \mathbb{R}^4 , making it $M' \equiv \mathbb{R}^4 - \{p\}$, which splits $\gamma(\lambda)$ into two geodesics $\gamma'(\lambda)$ and $\gamma''(\lambda)$, whose affine parameter λ has ranges (the domains of the curve maps γ' and γ'') $(-\infty, \lambda_p)$ and (λ_p, ∞) , respectively. We refer to both $\gamma'(\lambda)$ and $\gamma''(\lambda)$ incomplete geodesics. Generally speaking, an inextendible geodesic in (M, g_{ab}) is called an **incomplete geodesic** if the range of its affine parameter is not $(-\infty, \infty)$. The existence of an incomplete geodesic, to some extent, can be regarded as the sign of some region in spacetime being eliminated (and thus there is a “hole”). Hence, one may consider a definition as follows: “if there exists one (or more than one) incomplete geodesic in spacetime, then we call it a singular spacetime.” However, this definition has a serious flaw, namely the “scope of

attack” is overly broad. A spacetime that should not have been singular, if we remove a point by hand, would be a singular spacetime according to the preceding definition, which is not what we want. One way to overcome this flaw is to add a restriction in the definition: the spacetime we consider must be inextendible, i.e., it cannot be enlarged by adding some points to it.⁸ A spacetime with some points removed artificially is not inextendible, and hence does not meet this definition. Then we inspect the above definition from the perspective of whether it is physically singular. If there exists an incomplete timelike geodesic in an inextendible spacetime, physically it is indeed quite singular: the freely falling observer it represents will actually vanish in the spacetime within a finite time (according to its own standard clock) or not even have existed a finite amount of time earlier! Similarly, an incomplete null geodesic is also physically singular, since it represents the world line of a photon. However, a spacelike geodesic is not the world line of any particle, and so there is no reason to consider a spacetime which only has incomplete spacelike geodesics as physically singular. Hence, we take the following definition [see Hawking and Ellis (1973)]:

Definition 1 If there exists one (or more than one) incomplete timelike or null geodesic in an inextendible spacetime, we say that it is a **singular spacetime**, or it has a **spacetime singularity**.

However, Definition 1 still has drawbacks. For instance, there exists such a spacetime [see Geroch (1968)] which has no incomplete geodesics, but has a bizarre non-geodesic timelike curve (which has been maximally extended) whose arc length is finite and 4-acceleration (magnitude) is bounded. This indicates that the observer in a spaceship traveling along the curve will vanish in the spacetime after a finite time! (The finite arc length and bounded 4-acceleration assures that the spaceship can finish this curve with a finite amount of fuel, and a spaceship like this exists in principle.) Such a spacetime is singular enough to be called a singular spacetime, but unfortunately it is not according to Definition 1. This indicates that this definition has a drawback that the “scope of attack” is too narrow. Another drawback of Definition 1 is that the intuitive statement that the spacetime has a “hole” does not always meet the existence of an incomplete geodesics. For example, there exists such a geodesically incomplete spacetime (which contains an incomplete timelike, null or spacelike geodesic) whose background manifold is compact, and hence has no “holes” (according to Theorem 1.3.9, any point sequence in a compact manifold has an accumulation point, and thus the manifold has no “holes”), see Wald (1984). Although Definition 1 has these drawbacks, it may still be considered as the first choice of the definition of a singularity. The proof of the Penrose-Hawking singularity theorems used exactly this definition (Appendix E of Volume II provides a brief introduction of singularity theorems). Later we will see that in the maximally extended Schwarzschild spacetime there still exist many incomplete timelike and null geodesics (whose existence is related to the elimination of $r = 0$). Therefore,

⁸ The precise mathematical definition is: a spacetime (M, g_{ab}) is said to be inextendible if there does not exist a spacetime (M', g'_{ab}) such that there exists an isometry between the proper subsets of (M, g_{ab}) and (M', g'_{ab}) .

Schwarzschild spacetime is a singular spacetime, which has a spacetime singularity at $r = 0$ (and thus the points at $r = 0$ do not belong to Schwarzschild spacetime).

Many singular spacetimes have “curvature divergence” when an incomplete geodesic is approaching the singularity. The curvature is a tensor, whose components depend on the basis. The components of an ordinary tensor will also diverge in a bad basis, and therefore when talking about curvature divergence one needs to first give it a clear and valid definition. First, we may consider all kinds of scalars constructed by R_{abc}^d , g_{ab} and ∇_a [such as R , $R_{ab}R^{ab}$, $R_{abcd}R^{abcd}$, $R_{abcd}R^{cdef}R_{ef}^{ab}$, $(\nabla^c R^{ab})\nabla_c R_{ab}$] and their polynomials. If one of these quantities is divergent along an incomplete geodesic, then we say there exists an **s.p. curvature singularity**, where s.p. is the abbreviation for scalar polynomial. However, there also exists such a spacetime, whose scalar polynomials all vanish while $R_{abc}^d \neq 0$ (which is similar to the fact that the self-contraction of a null vector vanishes while the vector itself is not). Hence, we should also consider an alternative definition for curvature divergence: if at least one component of R_{abc}^d and its covariant derivatives in any frame parallelly transported along the geodesics is divergent, we say that the spacetime has a **p.p. curvature singularity**, where p.p. stands for parallelly propagated basis. Note that an s.p. curvature singularity contains a p.p. curvature singularity, but not vice versa. Once we find at least one incomplete timelike or null geodesic in the spacetime, we can say that the spacetime is singular. Then we inspect if these incomplete geodesics have curvature divergence, which has three possibilities: ① there is an s.p. singularity; ② there is a p.p. singularity but no s.p. singularity; ③ there is no curvature singularity (no curvature divergence). Taub-NUT spacetime is an example of a singular spacetime without curvature singularity [see Hawking and Ellis (1973) Sect. 5.8 and p. 261]. On the other hand, although some spacetimes have curvature divergence, they only diverge when “approaching infinity”, which should not be regarded as singular spacetimes. Thus, it is inappropriate to define spacetime singularity in terms of only curvature divergence but not geodesically incompleteness.

9.4.2 Coordinate Singularities of Rindler Metrics

If you can find a coordinate system such that the components of the Schwarzschild metric in this system behave ordinarily at $r = 2M$, you can claim that $r = 2M$ is only a coordinate singularity. This is a sufficient condition for determining a coordinate singularity. Unfortunately, finding this kind of “good” coordinate system in general is not easy, and is in no way guaranteed. Luckily, the singularity of the Schwarzschild metric at $r = 2M$ involves only the first two dimensions in the total 4-dimensional line element, and finding a “good” coordinates system in a 2-dimensional spacetime is way easier than doing that in a 4-dimensional spacetime. In this section we will first introduce a simple but heuristic example. Consider the 2-dimensional **Rindler spacetime**, whose metric has the following line element expression in the coordinates system $\{t, x\}$:

$$ds^2 = -x^2 dt^2 + dx^2. \quad (9.4.2)$$

The determinant of the metric components, $g = -x^2$, vanishes at $x = 0$, and hence the matrix $(g_{\mu\nu})$ has no inverse (is degenerate), which means $g_{\mu\nu}$ has a singularity at $x = 0$. We want to show that this is a coordinate singularity. First, note that the range of x should not include $x = 0$. There is a basic stipulation in relativity, namely the background manifold has to be a connected manifold. Therefore, the range of x can either be $x > 0$ or $x < 0$, but not the union of both. Without loss of generality, we take $x > 0$, i.e., we restrict the range of t, x to be

$$-\infty < t < \infty, \quad 0 < x < \infty. \quad (9.4.3)$$

The approach of finding a “good” coordinate system for determining the singularity at $x = 0$ as a coordinate singularity is based on the following fact: each point in a 2-dimensional spacetime has only two null directions (while in 4-dimensional spacetime there are infinitely many), and hence (locally) there are only two null geodesics passing through each point, which sorts the null geodesics in the entire spacetime into two families. If we find that a null geodesic is incomplete, then we should suspect that certain regions have been eliminated from the given spacetime. If one can show that these eliminated regions can be mended, i.e., the given spacetime can be extended, and $x = 0$ is a point in the extended spacetime, then one can claim that the singularity at $x = 0$ is only a coordinate singularity. Here is how it works:

Suppose $\eta(\lambda)$ is a null geodesic in Rindler spacetime, with λ as the affine parameter, then its tangent vector

$$(\partial/\partial\lambda)^a = (\partial/\partial t)^a dt/d\lambda + (\partial/\partial x)^a dx/d\lambda$$

satisfies

$$0 = g_{ab}(\partial/\partial\lambda)^a(\partial/\partial\lambda)^b = g_{00}(dt/d\lambda)^2 + g_{11}(dx/d\lambda)^2 = -x^2(dt/d\lambda)^2 + (dx/d\lambda)^2.$$

Thus, for $\eta(\lambda)$ we have

$$dt/dx = \pm 1/x, \quad t = \pm \ln x + c \quad (c \text{ is the constant of integration}), \quad (9.4.4)$$

where the positive sign and negative sign represents the “ingoing” family and “outgoing” family of null geodesics, respectively (the “ingoing” and “outgoing” here are introduced simply for the sake of convenience, one can choose either family as ingoing, and the other as outgoing), and different values of c correspond to different geodesics in the same family. Hence, $t + \ln x$ and $t - \ln x$ are constants on each “ingoing” and “outgoing” null geodesic. Define coordinates v and u as follows:

$$v := t + \ln x, \quad u := t - \ln x. \quad (9.4.5)$$

Then v is a constant on each “ingoing” null geodesic, and u is a constant on each “outgoing” null geodesic. It follows from (9.4.5) that

$$t = \frac{1}{2}(v + u), \quad x = e^{\frac{1}{2}(v-u)}, \quad (9.4.6)$$

and plugging these into (9.4.2) after differentiating them yields

$$ds^2 = -e^{v-u} dv du. \quad (9.4.7)$$

Thus, $0 = g_{vv} = g_{ab}(\partial/\partial v)^a(\partial/\partial v)^b$, which indicates that the basis vector $(\partial/\partial v)^a$ is a null vector. Similarly, $(\partial/\partial u)^a$ is also a null vector. Therefore, we refer to v and u as null coordinates. The ranges of the coordinates t and x [see (9.4.3)] correspond to the following ranges of v and u (see Fig. 9.11):

$$-\infty < v < \infty, \quad -\infty < u < \infty. \quad (9.4.8)$$

This seems to suggest that all the null geodesics are complete, but actually it does not since v and u are not affine parameters. The affine parameters can be obtained by means of the timelike Killing vector field $(\partial/\partial t)^a$. According to Theorem 4.3.3, the E defined as follows is a constant along any null geodesic $\eta(\lambda)$

$$E := -g_{ab}(\partial/\partial t)^a(\partial/\partial t)^b = -g_{00}dt/d\lambda = x^2 dt/d\lambda, \quad (9.4.9)$$

where λ is the affine parameter. Noticing that u is a constant on any “outgoing” null-geodesic, we can plug (9.4.6) into (9.4.9) and get

$$d\lambda = \frac{e^{-u}}{2E} e^v dv, \quad \lambda = \frac{e^{-u}}{2E} \int e^v dv = \frac{e^{-u}}{2E} e^v + c_1, \quad c_1 = \text{constant}. \quad (9.4.10)$$

Define

$$V := e^v. \quad (9.4.11)$$

Since $e^{-u}/2E$ and c_1 are constants, and λ is an affine parameter, (9.4.10) indicates that $V \equiv e^v$ is also an affine parameter of the “outgoing” null geodesics (see Theorem 3.3.3). From (9.4.8) and $V \equiv e^v$ we can see that the range of V is $(0, \infty)$, and thus the “outgoing” null geodesics are incomplete. Similarly, for “ingoing” null geodesics,

$$U := -e^{-u} \quad (9.4.12)$$

is an affine parameter. From (9.4.8) and (9.4.12) we know that the range of U is $(-\infty, 0)$, and thus the “ingoing” null geodesics are incomplete also. Does this indicate that Rindler spacetime is a singular spacetime with spacetime singularity at $x = 0$? No, the key is that Rindler spacetime is not an inextendible spacetime, but is the result of eliminating certain regions from a larger spacetime. To confirm this conclusion, one can derive from (9.4.11) and (9.4.12) that

$$dV dU = e^{v-u} dv du, \quad (9.4.13)$$

and plugging into (9.4.7) yields

$$ds^2 = -dVdU. \quad (9.4.14)$$

The range of the new coordinates V and U :

$$0 < V < \infty, \quad -\infty < U < 0 \quad (9.4.15)$$

are derived from the range of the original coordinates x , i.e., $0 < x < \infty$. However, now it is not necessary to stick to this range, since it follows from (9.4.14) that the only nonvanishing component in the coordinate system $\{V, U\}$ of the metric, $g_{VU} = -1/2$, behaves quite normally. Even if V, U take values exceeding the range of (9.4.15), the line element (9.4.14) still behaves well, with no singularity at all. If we present (9.4.15) to you in the first place without mentioning the previous discussion, you would naturally consider that the range of V, U has no constraints, i.e., they can take any value within $(-\infty, +\infty)$. In this way, the extension of the domain of the Rindler metric is realized by introducing new coordinates V, U . $x = 0$ represents points in the extended domain (the positive semi-axis of the V -axis, see Fig. 9.12). The metric behaves normally at these points, just its components in the original coordinate system $\{t, x\}$ behave badly there. This is actually pretty natural, since $x = 0$ never belongs to the coordinates patch of the original coordinate system (it only “touches the edge” from the outside), the so-called singularity at $x = 0$ is nothing but applying the original coordinate system inappropriately outside the coordinate patch. Therefore, the singularity at $x = 0$ is only a coordinate singularity. If we further define coordinates T, X as follows:

$$T := \frac{V + U}{2}, \quad X := \frac{V - U}{2}, \quad (9.4.16)$$

then it follows from (9.4.14) that $ds^2 = -dT^2 + dX^2$. Thus, the Rindler metric is actually a flat metric,⁹ just that its true colors are concealed by the original coordinates t, x . The Rindler spacetime defined in (9.4.2) is nothing but a sub-spacetime of the 2-dimensional Minkowski spacetime [a quadrant defined by (9.4.15), see region R in Fig. 9.12]. The Minkowski spacetime in Fig. 9.12 is the maximal extension of the Rindler spacetime in Fig. 9.11. It follows from $x^2 = e^{v-u} = -VU$ that both of the two lines $V = 0$ and $U = 0$ in Fig. 9.12 correspond to $x = 0$, which is exactly a specific manifestation of “ $x = 0$ does not belong to the coordinate patch of the original coordinate system (it only “touches the edge” from the outside)”. Although the two families of null geodesics in Fig. 9.11 appear differently from those in region R of Fig. 9.12, they are essentially the same. This again indicates that, even for the same spacetime, the spacetime diagram can vary widely due to different choices of the coordinate system.

⁹ It only differs from the Minkowski metric up to a diffeomorphism, and thus they are equivalent (they have the same geometry, see Sect. 8.10.2).

Fig. 9.11 The behavior of the “ingoing” family (1) and outgoing family (2) of null geodesics in the 2-dimensional Rindler spacetime in the coordinate system $\{t, x\}$

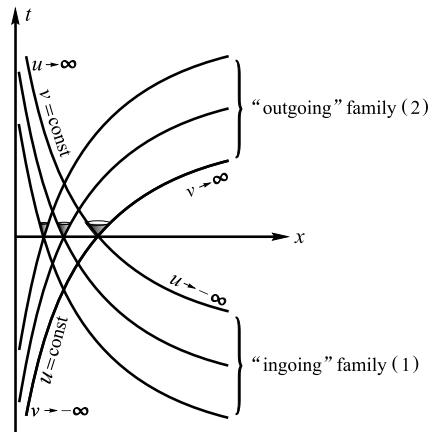
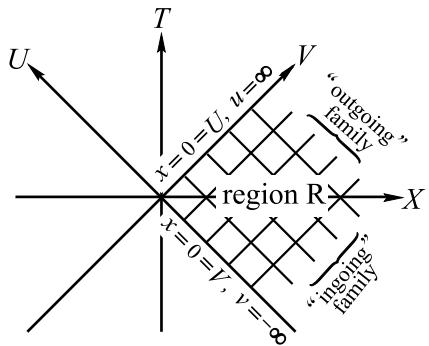


Fig. 9.12 2-dimensional Rindler spacetime is a sub-spacetime (region R) of the 2-dimensional Minkowski spacetime



9.4.3 The Kruskal Extension of Schwarzschild Spacetimes

As a differential equation, Einstein’s equation is local (one can talk about a differential equation at any given point and its neighborhood on the manifold). Each solution of the equation represents a metric. As for what manifold is the metric defined on (this is a global problem), one can only discuss it after solving the equation. Take the original Schwarzschild line element (9.4.1) as an example. We have pointed out that this line element has singularities at $r = 0$ and $r = 2M$. Since the background manifold has to be connected, the range of r can either be $r > 2M$ or $r < 2M$, but cannot be their union. We may take $r > 2M$, and then try to show that $r = 2M$ is a coordinate singularity. The way of proving this is quite similar to that of the Rindler case. The Rindler line element (9.4.2) is not only 2-dimensional, but also has a timelike Killing vector field $(\partial/\partial t)^a$, i.e., the metric components do not contain t , which greatly simplifies the task of finding a “good” coordinate system. We may summarize the way of accomplishing this task as the following procedure:

$$ds^2 = -x^2 dt^2 + dx^2 = x^2(-dt^2 + x^{-2}dx^2).$$

Define a function $x_*(x)$ such that $dx_* = x^{-1}dx$, then $ds^2 = x^2(-dt^2 + dx_*^2)$. Let $v := t + x_*$, $u := t - x_*$, i.e., $t = (v + u)/2$, $x_* = (v - u)/2$, then $-dt^2 + dx_*^2 = -dvdu$. Hence

$$ds^2 = -x^2 dvdu = -e^{v-u} dvdu = -dV dU,$$

where $V = e^v$, $U := -e^{-u}$. The Schwarzschild metric also has a Killing vector field $(\partial/\partial t)^a$, or equivalently, the coefficients of the first two dimensions of its line element (9.4.1) (denoted by $d\hat{s}^2$) do not contain t , and so the previous procedure is still applicable:

$$\begin{aligned} d\hat{s}^2 &= -(1 - 2M/r)dt^2 + (1 - 2M/r)^{-1}dr^2 \\ &= (1 - 2M/r)[-dt^2 + (1 - 2M/r)^{-2}dr^2] = (1 - 2M/r)(-dt^2 + dr_*^2), \end{aligned} \quad (9.4.17)$$

where

$$dr_* := (1 - 2M/r)^{-1}dr. \quad (9.4.18)$$

Take

$$r_* := r + 2M \ln \left(\frac{r}{2M} - 1 \right), \quad (9.4.19)$$

which is the tortoise coordinate r_* in (8.9.1). Let

$$v := t + r_*, \quad u := t - r_* \quad \text{or} \quad t = \frac{v+u}{2}, \quad r_* = \frac{v-u}{2}, \quad (9.4.20)$$

then the ranges of v and u are

$$-\infty < v, u < \infty. \quad (9.4.21)$$

It follows from (9.4.20) that $-dt^2 + dr_*^2 = -dvdu$, and hence

$$d\hat{s}^2 = -(1 - 2M/r)dvdu. \quad (9.4.22)$$

Let

$$V := e^{\beta v}, \quad U := -e^{-\beta u} \quad (\beta \text{ is an undetermined constant}), \quad (9.4.23)$$

then the ranges of V and U are

$$0 < V < \infty, \quad -\infty < U < 0. \quad (9.4.24)$$

Also

$$dv du = \beta^{-2} e^{\beta(u-v)} dV dU,$$

and hence

$$d\hat{s}^2 = -\beta^{-2} \left(\frac{r-2M}{r} \right) e^{\beta(u-v)} dV dU.$$

The factor $e^{\beta(u-v)}$ on the right-hand side of the above equation can be expressed using (9.4.20) as $e^{\beta(u-v)} = e^{-2\beta r_*}$. Using (9.4.19) to express $-2\beta r_*$, we can organize that

$$e^{\beta(u-v)} = e^{-2\beta r} \left(\frac{2M}{r-2M} \right)^{4\beta M}.$$

Hence,

$$d\hat{s}^2 = -\beta^{-2} \left(\frac{r-2M}{r} \right) e^{-2\beta r} \left(\frac{2M}{r-2M} \right)^{4\beta M} dV dU.$$

The cases where the above equation may be singular are $r = 0$ and $r - 2M = 0$, in which the latter can be eliminated by choosing

$$\beta = \frac{1}{4M} \quad (9.4.25)$$

as

$$d\hat{s}^2 = -\beta^{-2} \frac{2M}{r} e^{-2\beta r} dV dU = -\frac{32M^3}{r} e^{-r/2M} dV dU. \quad (9.4.26)$$

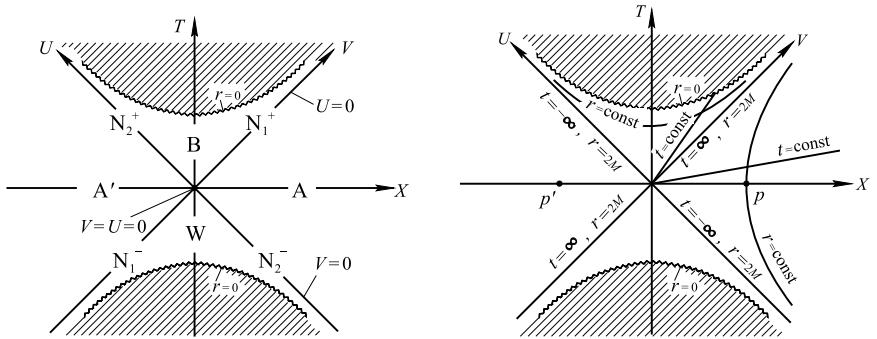
This equation indicates that the metric components are no longer singular at $r = 2M$, and hence the range of V, U can be extended to the regions where $V \leq 0$ and $U \geq 0$. Unlike the Rindler case, (9.4.26) indicates that $r = 0$ is still a singularity, and thus the range of r is constrained to $r > 0$. Thus, the values of V and U are in no way arbitrary; together they must satisfy the condition $r > 0$. Also let

$$T := \frac{1}{2}(V + U), \quad X := \frac{1}{2}(V - U), \quad (9.4.27)$$

and complete it with the other two dimensions, then we obtain the expression for the line element of the Schwarzschild metric in the Kruskal coordinate system $\{T, X, \theta, \phi\}$:

$$ds^2 = \frac{32M^3}{r} e^{-r/2M} (-dT^2 + dX^2) + r^2(d\theta^2 + \sin^2 \theta d\varphi^2). \quad (9.4.28)$$

The above equation indicates that the Schwarzschild metric can be defined on a manifold much larger than the original domain ($r > 2M$). Generally speaking, a spacetime $(\tilde{M}, \tilde{g}_{ab})$ is called an **extension** of a spacetime (M, g_{ab}) if $M \subset \tilde{M}$ and $\tilde{g}_{ab}|_p = g_{ab}|_p, \forall p \in M$. The extension of the original Schwarzschild spacetime we



(a) A and A' represent two asymptotically flat regions with no causal relation. B is the black hole region. W is the white hole region. N_1^+ and N_2^+ are the event horizons of the black hole. The zigzag curves represent the singularity (which does not belong to the spacetime).

(b) $r = \text{constant}$ on the hyperbolas, $t = \text{constant}$ on a line passing through the origin, $r = 2M$, $t = \pm\infty$ on the two lines tilted at 45° .

Fig. 9.13 The maximal Kruskal extension of Schwarzschild spacetime

obtained just now is called the **Kruskal extension** [Kruskal (1960)]. In this extension, the coordinates T, X can take all the values allowed by $r > 0$. The r in the line element (9.4.28) should be regarded as a function of the coordinates T and X , which is defined as follows (it is not difficult to prove this from the relation of the old and new coordinates):

$$\left(\frac{r}{2M} - 1\right) e^{r/2M} = X^2 - T^2. \quad (9.4.29)$$

Due to the spherical symmetry, one can sketch the spacetime diagram with only the first two dimensions (see Fig. 9.13), and by imagining each point in the diagram as an S^2 (2-dimensional sphere) yields the 4-dimensional spacetime. The factor $-dT^2 + dX^2$ in (9.4.28) indicates that in the 2-dimensional Schwarzschild spacetime diagram with T and X as the coordinate axes, the (radial) null curves are all lines with slope ± 1 . This will bring huge convenience to our discussion.

From (9.4.29) we can see that $r = \text{constant}$ corresponds to $X^2 - T^2 = \text{constant}$, i.e., a hyperbola in the TX -plane (a pair of 45° lines when $r = 2M$), which becomes a circular hyperboloid with the other two dimensions no longer suppressed (a hypersurface in the 4-dimensional manifold). There are two important special cases:

(1) $r = 0$ corresponds to $X^2 - T^2 = -1$. Thus, the bound of the Kruskal extension, $r > 0$, can be expressed in terms of the coordinates as

$$X^2 - T^2 > -1. \quad (9.4.30)$$

It is not difficult to show that any radial null or timelike geodesic with $r \rightarrow 0$ is incomplete. By calculation one also finds that the value of the scalar field $R_{abcd}R^{abcd}$ approaches ∞ when $r \rightarrow 0$ along these geodesics (which is obviously distinct from the fact that $R_{abcd}R^{abcd}$ approaches a finite value when $r \rightarrow 2M$), and thus there exists an s.p. curvature singularity. This implies that the spacetime cannot be

extended to $r = 0$ and beyond ($r < 0$), which means $r = 0$ is a spacetime singularity, and the Kruskal extension is the **maximal extension** of Schwarzschild spacetime. The shadow region in Fig. 9.13a does not belong to the extended spacetime. The two zigzag hyperbolas stand for the spacetime singularity at $r = 0$ (the singularity in 4-dimensional spacetime is not a point). The spacetime domain is an open subset of \mathbb{R}^2 , which is homeomorphic to \mathbb{R}^2 , while the topological structure of a 2-dimensional sphere is S^2 , and thus the topological structure of the maximally extended 4-dimensional Schwarzschild spacetime is $\mathbb{R}^2 \times S^2$. Note that each point outside the shadow region in Fig. 9.13 represents an S^2 , and two different points stand for two different S^2 . Especially, for example, point p and p' stand for neither the same S^2 , nor two points on the same S^2 , but each corresponds to one S^2 .

(2) $r = 2M$ corresponds to $X^2 - T^2 = 0$, i.e., $T = \pm X$, which represents two lines tilted at 45° passing through the origin in the 2-dimensional diagram (N_1 and N_2 in Fig. 9.13a). In the 4-dimensional spacetime they are two 3-dimensional surfaces (null hypersurfaces), which divide the spacetime into four open regions: region A is characterized by $X > 0$ and $X^2 > T^2$ (i.e., $V > 0, U < 0$). It follows from (9.4.29) that the region corresponds to the spacetime region of $r > 2M$; this is the coordinate patch of the original coordinate system $\{t, r\}$, which is also the “base area” where the Kruskal extension starts from. Treat Fig. 9.13a as the \tilde{M} in the above mentioned general definition of an extended spacetime, whose metric \tilde{g}_{ab} is described by the line element (9.4.28); treat the region A as the manifold M , whose metric g_{ab} is described by the line element (9.4.1). The line element (9.4.28) is nothing but the result of applying a coordinate transformation on (9.4.1); they represent the same metric field g_{ab} in the region A, and hence $(\tilde{M}, \tilde{g}_{ab})$ is indeed an extension of (A, g_{ab}) (the original Schwarzschild spacetime). All of B, W and A' are the outcome of the extension starting from A. The boundary points of A satisfy $V = 0$ and $U = 0$, while from (9.4.20) and the definition of V and U we get $t = 2M[\ln V - \ln(-U)]$. Thus, the points of A have $t \rightarrow \pm\infty$ when approaching the boundary (see Fig. 9.13b), which indicates that t is not defined on the two tilted lines. This is exactly the reason why the Schwarzschild line element (9.4.1) behaves singularly at $r = 2M$ (the two lines do not belong to the coordinate patch of the coordinate system $\{t, r\}$, but only “touches the edge” from the outside).

The coordinate t is not defined yet in the three regions B, W and A'. Recall the relations between the coordinates V , U and t, r_* in region A:

$$V = \exp[(r_* + t)/4M], \quad U = -\exp[(r_* - t)/4M]. \quad (9.4.31)$$

Reversely, we can define the t coordinates in the other three regions in terms of V , U using the following relations:

Region B	$V = \exp[(r_* + t)/4M], \quad U = \exp[(r_* - t)/4M],$
Region W	$V = -\exp[(r_* + t)/4M], \quad U = -\exp[(r_* - t)/4M],$ (9.4.31')
Region A'	$V = -\exp[(r_* + t)/4M], \quad U = \exp[(r_* - t)/4M],$

where

$$r_* \equiv r + 2M \ln |r/2M - 1|. \quad (9.4.32)$$

Applying the line element (9.4.28) to regions B, W, A' and rewriting the line element in terms of t, r by means of (9.4.31') and (9.4.32), we still get (9.4.1), where the range of r for regions B, W is $0 < r < 2M$, and for A' is $r > 2M$. Thus, the metric in A, A' and B, W are, respectively, the Schwarzschild line element (9.4.1) restricted to $r > 2M$ and $0 < r < 2M$. The relations between the coordinate T, X and t, r in these 4 regions are as follows:

$$\begin{aligned} \text{Region A} \quad T &= (r/2M - 1)^{1/2} e^{r/4M} \sinh(t/4M), \\ X &= (r/2M - 1)^{1/2} e^{r/4M} \cosh(t/4M), \end{aligned} \quad (9.4.33)$$

$$\begin{aligned} \text{Region B} \quad T &= (1 - r/2M)^{1/2} e^{r/4M} \cosh(t/4M), \\ X &= (1 - r/2M)^{1/2} e^{r/4M} \sinh(t/4M), \end{aligned} \quad (9.4.34)$$

$$\begin{aligned} \text{Region W} \quad T &= -(1 - r/2M)^{1/2} e^{r/4M} \cosh(t/4M), \\ X &= -(1 - r/2M)^{1/2} e^{r/4M} \sinh(t/4M), \end{aligned} \quad (9.4.35)$$

$$\begin{aligned} \text{Region A}' \quad T &= -(r/2M - 1)^{1/2} e^{r/4M} \sinh(t/4M), \\ X &= -(r/2M - 1)^{1/2} e^{r/4M} \cosh(t/4M). \end{aligned} \quad (9.4.36)$$

The inverse transformations are

$$\text{Regions A, B, W, A}' \quad (r/2M - 1)e^{r/2M} = X^2 - T^2, \quad (9.4.37)$$

$$\text{Regions A, A}' \quad t/2M = 2 \tanh^{-1}(T/X), \quad (9.4.38)$$

$$\text{Regions B, W} \quad t/2M = 2 \tanh^{-1}(X/T). \quad (9.4.39)$$

We have mentioned at the beginning of this subsection that according to (9.4.1), on the one hand we cannot take $r = 2M$, while on the other hand we cannot take both $r > 2M$ and $0 < r < 2M$ (otherwise it would be disconnected). However, things will be different now that we have the Kruskal extension. This extension indicates that the Schwarzschild metric is defined on regions A, B and their intersection N_1^+ (on which $r = 2M, t = \infty$), and $A \cup N_1^+ \cup B$ is a connected manifold. An “ingoing” (r keeps decreasing), future-directed null curve starting from any point in A will inevitably cross N_1^+ and enter B. (However, a timelike curve can go to infinity with N_1^+ as the asymptote). In contrast, for a future-directed timelike or null curve starting from any point in B it will be impossible to cross N_1^+ and enter A, and its end can only be falling into the singularity. (The singularity does not belong to the spacetime. The precise meaning of “falling into the singularity” is that the r of this world line becomes smaller and smaller, and approaches zero. For a timelike geodesic, falling into the singularity means that the freely falling observer it represents vanishes from the spacetime when the proper time reaches a certain value, which is indeed incredibly singular.) This indicates that N_1^+ is a “one-way membrane” with no way out. Any

object (including a photon) in the region A can never return to A (but can only fall into the singularity) once it enters the region B. Therefore, the region B is called a **black hole**, and N_1^+ is called the **event horizon**. Considering that each point in Fig. 9.13 represents a 2-dimensional sphere, and thus the black hole is a 4-dimensional spacetime region, while the event horizon is a (3-dimensional) null hypersurface (the proof for the event horizon being a null hypersurface is left as Exercise 9.11, see the hint therein). The region A' is characterized by $X < 0$ and $X^2 > T^2$, and it also has $r > 2M$. In fact, it has exactly the same properties as the region A, including that its relationship with the black hole B is similar to the relationship between A and B, and hence N_2^+ is the event horizon of A' . However, A and A' do not have any causal relation: any timelike or null curve starting from A cannot enter A' and vice versa. In this sense, people also often refer to A and A' as two (independent) “universes”. The region W is characterized by $T < 0$ and $X^2 < T^2$, and it also has $r < 2M$. W and A (or A') are only divided by a “membrane”, which is the null hypersurface N_2^- (or N_1^-). Both N_2^- and N_1^- are “one-way membranes” with no way out. Any future-directed timelike or null curve in W will cross N_2^- (or N_1^-) and enter A (or A'). Since B is called a black hole, W is naturally called a **white hole**.

The above discussion is about the maximal extension of Schwarzschild spacetime obtained under the premise that the entire spacetime is a vacuum. Although this extension includes some tempting terminologies such as black hole, white hole, event horizon and the two identical “universes”, the physical existence (authenticity) of it deserves additional discussions. From the perspective of the initial value problem, the chance for this entire spacetime to exist is very small, while part of it (including part of A, B and the event horizon in between) is very meaningful, see Sect. 9.4.6 for details.

At the end of this subsection, we would like to discuss the Killing vector fields of the maximally extended Schwarzschild spacetime. Before the extension, the spacetime has 4 independent Killing vector fields, in which 3 of them reflect the spherical symmetry, see the $\xi_1^a, \xi_2^a, \xi_3^a$ in Sect. 8.2; the fourth one reflects the staticity, namely $\xi^a = (\partial/\partial t)^a$. For the maximally extended Schwarzschild spacetime, $\xi_1^a, \xi_2^a, \xi_3^a$ still reflect the spherical symmetry. Since the t coordinate is defined in all for regions A, A' , B, W, and the line elements in all regions written in terms of t, r are the original Schwarzschild form, the $\xi^a = (\partial/\partial t)^a$ in each region is still a Killing field. Note that in B and W ξ^a is not timelike but spacelike, since it follows from the line element (9.4.1) that $r < 2M$ leads to $g_{ab}(\partial/\partial t)^a(\partial/\partial t)^b > 0$. There does not exist other independent Killing vector fields besides $\xi_1^a, \xi_2^a, \xi_3^a$ and ξ^a , and hence B and W are not static spacetime regions. $(\partial/\partial t)^a$ is undefined on the null hypersurfaces N_1 and N_2 , since the coordinate t is undefined on it ($t = \pm\infty$). However, one can express ξ^a in A using the coordinate basis vectors $(\partial/\partial V)^a$ and $(\partial/\partial U)^a$ as

$$\xi^a = (\partial/\partial t)^a = \frac{1}{4M} [V(\partial/\partial V)^a - U(\partial/\partial U)^a]. \quad (9.4.40)$$

Since $(\partial/\partial V)^a$ and $(\partial/\partial U)^a$ are well-defined on N_1 and N_2 , one can define the vector field ξ^a on N_1 and N_2 using the above equation, and verify that it is a null Killing

vector field. Hence, on the whole manifold there is a fourth C^∞ Killing vector field ξ^a , which is orthogonal to the other 3 independent Killing vector fields. Thus, the symmetry of the maximally extended Schwarzschild spacetime is characterized by 4 Killing fields, in which three reflect the spherical symmetry and the fourth (i.e., ξ^a) is timelike in A and A', spacelike in B and W, and null on N_1 and N_2 . Thus, we can see the necessity of changing “static” to “Schwarzschild” in the original formulation of Birkhoff’s theorem “a spherically symmetric solution of the vacuum Einstein equation must be a static metric” (see Sect. 8.3.3): the Schwarzschild metric is not necessarily a static metric. From the geometric perspective, the essence of Birkhoff’s theorem is: if the metric satisfies the vacuum Einstein equation and has the three Killing vector fields reflecting the spherical symmetry, then it must have a fourth (additional, not preassigned) Killing vector field ξ^a , which can be timelike, spacelike, or even null, depending on where the spacetime point is located.

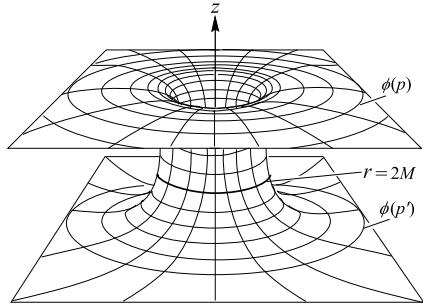
9.4.4 Surfaces of Infinite Redshift in Schwarzschild Spacetimes

Suppose the radius coordinates of static observers G and G' outside the event horizon are r and $r' (> r)$, respectively. G emits light toward G' . One can derive from (9.2.3) that the redshift $z \equiv (\lambda' - \lambda)/\lambda$. If r' is fixed, then z is a function of r satisfying $dz(r)/dr < 0$ and $\lim_{r \rightarrow 2M} z(r) = \infty$. Therefore, the hypersurface $r = 2M$ is also called the **surface of infinite redshift**. However, one should not say “the light emitted from the surface of infinite redshift will have an infinite redshift when it reaches G' ”, since any outgoing null geodesic emitted from the hypersurface $r = 2M$ (event horizon) can only lie on the horizon and can never reach G' .

The Schwarzschild spacetime (region A) has only one static reference frame (it has only one hypersurface orthogonal Killing vector field, namely ξ^a), but there are infinitely many stationary reference frames. This is because a linear combination of ξ^a and the spatial Killing field $(\partial/\partial\varphi)^a$, $\tilde{\xi}^a \equiv \xi^a + \beta(\partial/\partial\varphi)^a$ (where β is a constant) is also a Killing field, and so $\tilde{\xi}^a$ corresponds to a stationary reference frame in the region where it is timelike. Equation (9.2.2) can be applied to any stationary reference frame, where one just needs to interpret the ξ^a as $\tilde{\xi}^a$. The surface of infinite redshift corresponds to $-\tilde{\xi}^a\tilde{\xi}_a = 0$, and thus relies on the stationary reference frame. In fact, if one wants, one can even find a stationary reference frame that has a surface of infinite redshift for Minkowski spacetime. Since the static reference frame in Schwarzschild spacetime is unique, unless otherwise indicated, the surface of infinite redshift will refer to the surface $-\xi^a\xi_a = 0$ (which coincides with the event horizon), and the “redshift factor” will mean

$$\chi = (-\xi^a\xi_a)^{1/2} = (1 - 2M/r)^{1/2}.$$

Fig. 9.14 The embedding diagram of the maximally extended Schwarzschild spacetime ($T = 0$, one dimension suppressed)



9.4.5 Embedding Diagrams [Optional Reading]

Lots of literature, including textbooks and popular science books like to use embedding diagrams (see Fig. 9.14) to intuitively describe the Schwarzschild black hole. This subsection provides an introduction to embedding diagrams. To start with, we first discuss the simple case of the embedding diagram for a static spherically symmetric star. Equation (9.3.19) represents the metric inside a static spherically symmetric star, whose induced line element on any constant- t surface Σ_t reads

$$ds^2 = \left(1 - \frac{2m(r)}{r}\right)^{-1} dr^2 + r^2 (d\theta^2 + \sin^2 \theta d\varphi^2). \quad (9.4.41)$$

Let R be the radius of the star. If we let $m(r)$ take a constant value $M \equiv m(R)$ when $r \geq R$, then the above equation applies to both the inner and outer parts of the star. This is a curved line element. Due to the spherical symmetry, we can just consider the cross section with $\theta = \pi/2$ in Σ_t (denoted by S), whose induced line element is

$$ds^2 = \left(1 - \frac{2m(r)}{r}\right)^{-1} dr^2 + r^2 d\varphi^2. \quad (9.4.42)$$

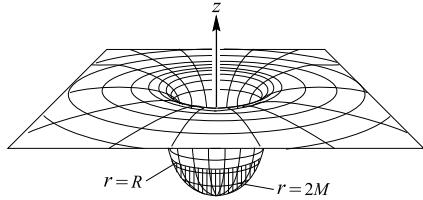
Let g_{ab} represent the metric corresponding to this line element, then (S, g_{ab}) is a 2-dimensional Riemannian space. To intuitively manifest its intrinsic warping, one can embed it into the one higher dimensional Euclidean space $(\mathbb{R}^3, \delta_{ab})$, i.e., consider the embedding $\phi : S \rightarrow \mathbb{R}^3$, and use the warping of $\phi[S]$ in \mathbb{R}^3 to intuitively reflect the intrinsic warping of (S, g_{ab}) . Figure 9.15 is the embedding diagram that embeds (S, g_{ab}) into $(\mathbb{R}^3, \delta_{ab})$. From this figure we can see that the further from the center of the star, the lesser the space warps, and as $r \rightarrow \infty$ it approaches flat space. However, how is this diagram drawn? Based on what principle can we draw a diagram with this kind of effect?

The line element expression of the 3-dimensional Euclidean metric δ_{ab} in a cylindrical coordinate system $\{z, r, \varphi\}$ reads

$$ds^2 = dz^2 + dr^2 + r^2 d\varphi^2. \quad (9.4.43)$$

Take a radial line segment on S . The difference between the values of r at its ends p and p' is dr (see Fig. 9.16 left). If we parallelly transport this segment to somewhere on S with a different value of r , then although the new segment has the same dr as the old one, the arc length is in general different [see (9.4.42)]. This is a significant manifestation of the intrinsic warping of (S, g_{ab}) . Let $q \equiv \phi(p)$ and $q' \equiv \phi(p')$. As long as we assure that the

Fig. 9.15 The embedding diagram of a static spherically symmetric star (one dimension suppressed)



line segments qq' and pp' have the same arc length when drawing the diagram, then the external warping of $\phi[S]$ in \mathbb{R}^3 reflects the above-mentioned intrinsic warping of (S, g_{ab}) . This is the principle of making the embedding diagram. Based on this one can find the equation of the surface $\phi[S]$, and then draw $\phi[S]$. As a hypersurface in \mathbb{R}^3 , the equation of $\phi[S]$ can be expressed as $f(z, r) = 0$ (the axial symmetry makes f to not depend on φ). This corresponds to a function of one variable, $z = z(r)$, which represents the dependence of the value of z on the value of r at an arbitrary point on $\phi[S]$. Hence, the arc length of any line segment on $\phi[S]$ is

$$ds^2 = dz^2 + dr^2 + r^2 d\varphi^2 = \{[dz(r)/dr]^2 + 1\}dr^2 + r^2 d\varphi^2. \quad (9.4.44)$$

Comparing this with (9.4.42) yields

$$\left[\frac{dz(r)}{dr} \right]^2 + 1 = \left(1 - \frac{2m(r)}{r} \right)^{-1}, \quad \text{i.e.,} \quad \frac{dz(r)}{dr} = \sqrt{\frac{2m(r)}{r - 2m(r)}}. \quad (9.4.45)$$

Stipulate $z(0) = 0$, then

$$z(r) = \int_0^r \sqrt{\frac{2m(r')}{r - 2m(r')}} dr' \quad (\text{for } 0 < r < \infty). \quad (9.4.46)$$

Since $m(r) = M$ for $r \geq R$, we have

$$z(r) = \sqrt{8M(r - 2M)} + C \quad (\text{for } r \geq R), \quad (9.4.47)$$

where

$$C \equiv -\sqrt{8M(R - 2M)} + \int_0^R \sqrt{\frac{2m(r)}{r - 2m(r)}} dr \quad (9.4.48)$$

is a constant. Although for points with $r < R$, $z(r)$ depends on the form of the function $m(r)$, $r > 2m(r)$ guarantees that $z(r)$ is monotonically increasing, and (9.4.46) indicates that $\phi[S]$ is a circular paraboloid when $r > R$. Hence, we have the embedding diagram shown in Fig. 9.15 (the $r < R$ part is only a qualitative sketch). Note that the background Euclidean space ($\mathbb{R}^3, \delta_{ab}$) is introduced by hand only for showing $\phi[S]$, and the points with actual physical meaning are only the points on $\phi[S]$. (Do not think there is something filled in the “hat”!)

Now it is not difficult to understand Fig. 9.14, which is actually the embedding of the “whole space” Σ_0 at $T = 0$ (at $t = 0$) in the Kruskal extension of the Schwarzschild solution (see Fig. 9.13). Σ_0 contains all the points (each represents an S^2) on the X -axis in Fig. 9.13. Following the above derivation we can see that the hypersurface $\phi(\Sigma_0)$ in the embedding diagram (one dimension is suppressed) is a circular paraboloid determined by the equation

$$z(r) = \pm \sqrt{8M(r - 2M)}. \quad (9.4.49)$$

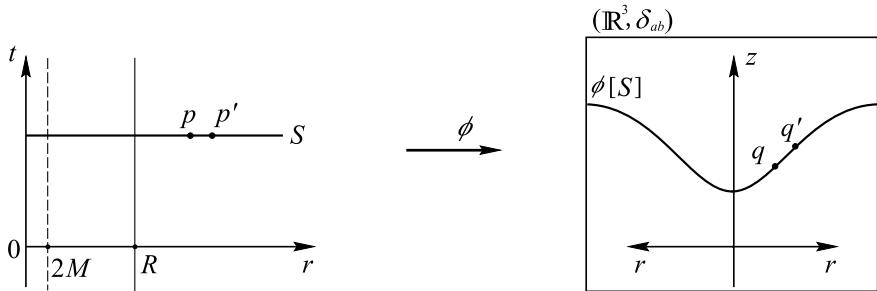


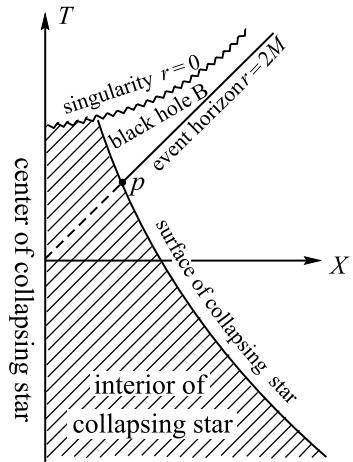
Fig. 9.16 Embedding a surface S in the spacetime of a static spherically symmetric star into $(\mathbb{R}^3, \delta_{ab})$

(Due to the asymptotic flatness, it develops from above and below into two surfaces that are approximately planes.) Since the value of r in any point of Σ_0 is greater than or equal to $2M$, there does not exist a point with $r < 2M$ in the embedding diagram. The whole “space” is divided by the circle formed by the points of $r = 2M$ (it is actually a sphere, called the **throat**) into upper and lower halves, which correspond respectively to the $X > 0$ and $X < 0$ parts on the X -axis in Fig. 9.13b. For instance, the points p and p' in Fig. 9.13b correspond respectively to the circles (spheres) $\phi(p)$ and $\phi(p')$ in Fig. 9.14. It is necessary to reiterate that only the circular paraboloid represents the “whole space” Σ_0 at $t = 0$, while the points outside the surface do not have any physical meaning.

9.4.6 The Gravitational Collapse of a Spherical Star and Schwarzschild Black Holes

As we have mentioned in Sect. 9.3.2, if a star in its late stage of evolution wants to maintain hydrostatic equilibrium in its interior [satisfying (9.3.17)], its mass must be less than the upper mass limit of a neutron star. If a star whose initial mass is greater than this upper bound cannot eject enough mass during its evolution and become a stable white dwarf or neutron star, then it cannot be stable at all but can only keep contracting until it becomes a black hole. According to Birkhoff’s theorem (see Sect. 8.3.3), the exterior of the star must have a Schwarzschild metric, and hence can be described by the spacetime diagram shown in Fig. 9.17. The non-shadow region in the diagram is identical to the corresponding part in Fig. 9.13, while the shadow region is described by the interior metric (non-vacuum solution to Einstein’s equation). Therefore, the spacetime of a collapsing star does not have the white hole region W at all, and it does not have the Region A' either, while the black hole region B and part of the region A in this case are of great significance. No matter how solid the matter constructing the star is, as long as the surface of the star crosses the event horizon, it has to keep contracting until the entire star is squashed into the singularity. The reason is simple: the world line of any point on the surface of the star must lie inside the light cone (must be timelike), and thus the angle between the T -axis and

Fig. 9.17 The late time collapse of a massive star described by the Kruskal coordinates. The vacuum Schwarzschild solution only applies to the exterior of the star's surface. The unshadowed region B represents the black hole caused by the collapse

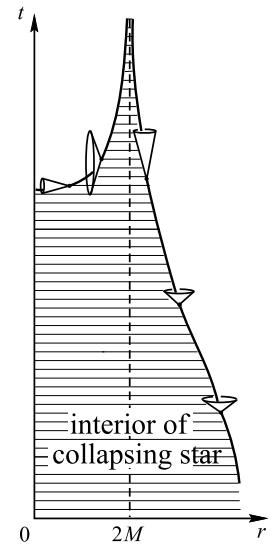


the world line must be smaller than 45° (note that the lines tilted at 45° in Figs. 9.13 and 9.17 represent radial null geodesics). The Schwarzschild coordinates can only cover the spacetime region of $r > 2M$ (or $0 < r < 2M$), and thus cannot manifest the whole process of a star in its late stage collapse into a black hole; particularly, it cannot manifest the most crucial step in the whole process—the surface of the star is contracted into the event horizon. If we want to describe the collapse of a star using the Schwarzschild coordinates, then we can only draw it as Fig. 9.18. Since the Schwarzschild coordinate t is undefined at $r = 2M$, this figure is actually just the result of combining two diagrams (representing $r > 2M$ and $0 < r < 2M$) together. The right part of the figure ($r > 2M$) may mislead people into thinking that the surface of the collapsing star is always outside the event horizon $r = 2M$. This kind of misunderstanding comes from confusing $t = \infty$ with “always” (readers who are familiar with Zeno’s paradox may notice the similarity between the coordinate time t and “Achilles time”). From Fig. 9.17 we can see that the intersection of the star’s surface and $r = 2M$ (see p in Fig. 9.17) corresponds to $t = \infty$. However, the proper time τ of an observer at a point p on the star’s surface has a finite value, and they will enter the black hole and fall into the singularity in a very short time $\Delta\tau$. (For a black hole with $M = M_\odot$, $\Delta\tau$ is approximately 2×10^{-5} s.)

The process of a star collapsing into a black hole can be represented more intuitively by another coordinate system—the ingoing Eddington-Finkelstein coordinate system $\{v, r, \theta, \varphi\}$. Although it cannot cover the maximally extended Schwarzschild spacetime like the Kruskal system, it can cover the regions A and B (unlike the Schwarzschild coordinate system which can only cover any one of the four regions). The r, θ, φ in this system are the same as the corresponding coordinates in the Schwarzschild system, and $v := t + r_*$. The line element of the first two dimensions in the Schwarzschild system

$$ds^2 = -(1 - 2M/r)dt^2 + (1 - 2M/r)^{-1}dr^2 \quad (9.4.50)$$

Fig. 9.18 The late time collapse of a massive star described by the Schwarzschild coordinates. Although not until $t \rightarrow \infty$ will the star's surface shrink to $r = 2M$, this does not indicate that it will always be outside the horizon, since the coordinate time approaching infinity does not represent “always”



in the ingoing Eddington-Finkelstein system becomes

$$ds^2 = -(1 - 2M/r)dv^2 + 2dvdr. \quad (9.4.51)$$

It follows from the equation above that the nonvanishing components $g_{vv} = -(1 - 2M/r)$, $g_{vr} = 1$ and the determinant $g = -1$ are all well-behaved at $r = 2M$, and hence $r = 2M$ is no longer a singularity. Also, considering that $v \in (-\infty, \infty)$ corresponds to $V \in (0, \infty)$, we can see that $\{v, r\}$ can cover the regions A and B in Fig. 9.13. $g_{rr} = 0$ and $g_{vv} = -(1 - 2M/r)$ also indicates that the coordinate basis vector $(\partial/\partial r)^a$ of the Eddington-Finkelstein system $\{v, r, \theta, \varphi\}$ is a null vector while $(\partial/\partial v)^a$ is a timelike (for region A) or spacelike (for region B) vector. Suppose $\eta(\lambda)$ is an arbitrary radial null geodesic in the regions A and B, then it follows from (9.4.51) that

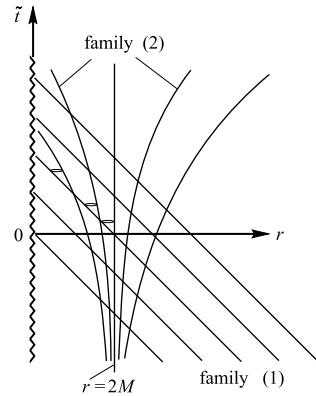
$$0 = -\left(1 - \frac{2M}{r}\right)\left(\frac{dv}{d\lambda}\right)^2 + 2\frac{dv}{d\lambda}\frac{dr}{d\lambda} = \frac{dv}{d\lambda}\left[-\left(1 - \frac{2M}{r}\right)\frac{dv}{d\lambda} + 2\frac{dr}{d\lambda}\right].$$

This indicates that the radial null geodesics can be classified into two families, which are characterized respectively by the following conditions:

$$(1) \quad \frac{dv}{d\lambda} = 0, \quad \text{i.e., } v = \text{constant}, \quad (9.4.52)$$

$$(2) \quad -\left(1 - \frac{2M}{r}\right)\frac{dv}{d\lambda} + 2\frac{dr}{d\lambda} = 0, \quad \text{and hence } \frac{dv}{dr} = \frac{2r}{r - 2M}. \quad (9.4.53)$$

Fig. 9.19 The behavior of the two families of null geodesics in 2-dimensional Schwarzschild spacetime in the coordinate system $\{\tilde{t}, r\}$



The first family of null geodesics are horizontal lines in the vr -diagram, which is not intuitive enough. Define $\tilde{t} := v - r$, then it follows from (9.4.51) that

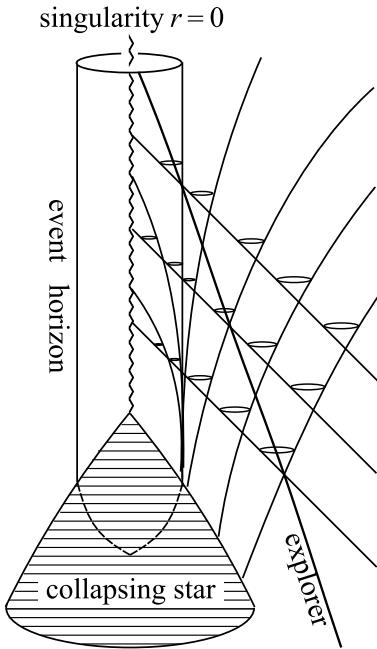
$$d\hat{s}^2 = -(1 - 2M/r)d\tilde{t}^2 + (4M/r)d\tilde{t}dr + (1 + 2M/r)dr^2. \quad (9.4.54)$$

The behaviors of the two families of null geodesics in the coordinate system $\{\tilde{t}, r\}$ are shown in Fig. 9.19: the equation for family (1) (incoming family) is $d\tilde{t}/dr = -1$, and thus it is a family of parallel lines with slope -1 ; the equation for family (2) (outgoing family) is

$$\frac{d\tilde{t}}{dr} = \frac{r + 2M}{r - 2M}.$$

The behavior of this family is rather special: all of the null geodesics are curves except for a vertical line ($r = 2M$); for a curve on the right of the vertical line, the value of r increases as the affine parameter λ increases (which is truly outgoing), while for a curve on the left of the vertical line, the value of r decreases as λ increases (which is actually ingoing, but belongs to the outgoing family). This oddity reflects an important property of the black hole: $r = 2M$ is the event horizon, and any photon inside the horizon ($r < 2M$) cannot cross the horizon and come out of the black hole (to $r > 2M$); their values of r can only keep decreasing to zero. Based on the two families of null geodesics one can easily draw the light cone at each point, which is helpful for analyzing the motion of a point mass, since the world line of a point mass is a timelike curve and the tangent vector at each point on the line must lie inside the light cone at this point. Thus, a point mass outside the event horizon can cross the horizon and enter the black hole, while once it enters there is no way out, and it can only fall in to the singularity. Revolving Fig. 9.19 with respect to the \tilde{t} -axis we get the 3-dimensional spacetime diagram (see Fig. 9.20), and then adding the world tube of the surface of the collapsing star (the cannon shape surface in Fig. 9.20) we can intuitively represent the exterior spacetime geometry of a star collapsing into a black hole. To facilitate understanding this, let us consider the

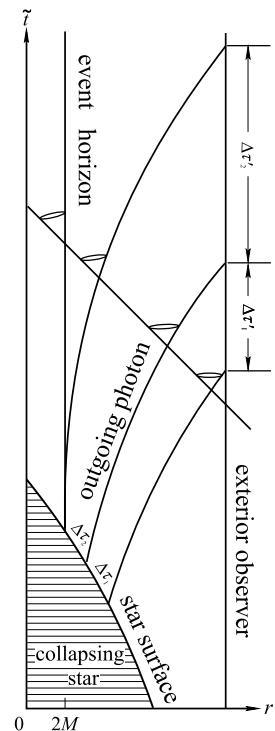
Fig. 9.20 The spacetime diagram of the star collapsing into a black hole in the system $\{\tilde{t}, r\}$. A black hole explorer will inevitably fall into the singularity if they do not turn around before reaching the horizon



following thought experiment (ignore the effects of the tidal force). Suppose you are an observer exploring a black hole on a spaceship with enough fuel. If you do not turn on the engine, the spaceship will fall freely, and will evidently cross the event horizon, enter the black hole and die at the singularity. If you turn around before reaching the horizon and go full steam ahead (namely let r increase before it reaches $2M$), you will be able to return safely and write an exploration report. However, if you go for one more step and reach the horizon (note that you will not feel anything special when your world line intersects the horizon), then this moment will become the regret of a lifetime, because from the light cone on the horizon we can see that you cannot return once you are at the horizon. You cannot even have a phone call wirelessly to your distant friend, since the “outgoing” photon on the horizon can only go upward vertically along the horizon (r remains equal to $2M$), see Fig. 9.20.

Now we will discuss the appearance of the collapsing star. Figure 9.21 shows the situation of a photon emitted from the surface of the star reaching an exterior static observer. Since the photon emitted from the surface of the star cannot reach outside the horizon, seemingly an exterior observer will find that the star gets smaller gradually and vanishes suddenly. However, if we look at Fig. 9.21 more carefully we will see that this is not the case. Since the world line of an outgoing photon outside the horizon will be more steep when it comes closer to the horizon, and becomes completely vertical on the horizon (lies on the horizon), the exterior observers will always (no matter how large its proper time τ' is) receive the light emitted from the surface of the star outside the horizon. They will observe that the star contracts slower

Fig. 9.21 An exterior observer in principle can always receive the light emitted from the star's surface. $\Delta\tau'_1 > \Delta\tau_1$ indicates that there exists redshift; $\Delta\tau_2 = \Delta\tau_1$ and $\Delta\tau'_2 > \Delta\tau'_1$ indicates that the redshift is getting stronger



and slower, and approaches a certain size,¹⁰ i.e., the radius of the star will approach $2M$ with a smaller and smaller speed and get “frozen” at this size. This phenomenon is also called the time dilation effect of the gravitational field. We have mentioned in Chap. 6 that when comparing the rates of clocks, first one needs to stipulate a specific method for “clock comparison”. In the situation of Fig. 9.21, the world lines of photons becomes the key of clock comparison: we stipulate the proper times of the world lines of, respectively, an observer on the star’s surface and an exterior observer set by the world lines of two neighboring radial photons, $\Delta\tau$ and $\Delta\tau'$, as the objects to compare. Calculation shows that (see Optional Reading 9.4.2) $\Delta\tau' > \Delta\tau$, and if $\Delta\tau_2 = \Delta\tau_1$ then $\Delta\tau'_2 > \Delta\tau'_1$, i.e., $\Delta\tau'/\Delta\tau$ increases as τ increases. Thus, the exterior observer views that a standard clock on the star’s surface is not only slower than his own clock, but also becomes slower and slower as it goes (note that this kind of “view” is the outcome of both the spacetime geometry and the method of clock comparison we just stipulated). Another manifestation of this effect of time dilation is the redshift. Regard two neighboring null geodesics as the world lines of two neighboring wave crests, then $\Delta\tau$ and $\Delta\tau'$ are the periods of the light waves measured by the observer

¹⁰ Later we will see that the light waves received by the exterior observers will have stronger and stronger redshift. Thus, only if we assume (theoretically) that the observer is sensitive to light of any wavelength and intensity can they observe this phenomenon.

on the star's surface and the exterior observer, respectively. $\Delta\tau' > \Delta\tau$ indicates that the light wave received by the exterior observer has a longer wavelength, i.e., it has a redshift, and $\Delta\tau'/\Delta\tau$ increasing with the increase of τ indicates that the redshift is getting stronger. However, this kind of redshift is different from the redshift between stationary observers which we discussed in Sect. 9.2.1, since the observer on the star's surface is not a stationary observer, see Optional Reading 9.4.2 for details.

[Optional Reading 9.4.1]

Beginners often feel confused by the fact that the spacetime diagrams of the same physical process in different coordinate systems can be so different. The essence of the problem is actually very simple: a coordinate system by definition is nothing but a map from an open set O of the manifold to an open set V of \mathbb{R}^n , and a spacetime diagram is a diagram in V . The same physical process can certainly have different spacetime diagrams in different coordinate systems. So we may say that a physical process is absolute, while a spacetime diagram is relative (since a coordinate system is involved). We have pointed this out when we first introduced spacetime diagrams in Sect. 6.1.5.

It follows from (9.4.54) that the coordinate basis vectors $(\partial/\partial\tilde{t})^a$ and $(\partial/\partial r)^a$ are not orthogonal. In Fig. 9.19, the \tilde{t} -axis and the r -axis are drawn to be orthogonal; this is because the spacetime diagram is a diagram in an open set V of \mathbb{R} , which does not reflect the spacetime metric, and thus does not reflect the orthogonality of vectors. All Fig. 9.19 tells us is: all the vertical lines have r as a constant, all the horizontal lines have \tilde{t} as a constant. Only in this way can the two families of null geodesics characterized by $d\tilde{t}/dr = -1$ and $d\tilde{r}/dr = (r + 2M)/(r - 2M)$ be represented as the two families of curves in the diagram.

[The End of Optional Reading 9.4.1]

[Optional Reading 9.4.2]

To simplify the discussion, we consider the simplest model of a star, i.e., a spherically symmetric star with a uniform density and no pressure (i.e., a dust cloud). Since the pressure gradient is zero, the world line of each point on the star's surface is a radial timelike geodesic. Figure 9.22 shows the situation of a photon emitted from an event p on the star's surface reaching an exterior observer (event p'). Z^a and \tilde{Z}^a represent the 4-velocities of a radial freely falling observer and a static observer at p , respectively; Z'^a represents the 4-velocity of an exterior static observer at p' . Suppose λ , $\tilde{\lambda}$ and λ' are the wavelengths of the same photon measured by Z^a , \tilde{Z}^a and Z'^a , respectively, then it follows from (9.4.55) that

$$\frac{\lambda'}{\tilde{\lambda}} = \frac{\chi'}{\chi}, \quad (9.4.55)$$

where

$$\chi \equiv (-\xi^a \xi_a)^{1/2}|_p = \left[1 - \frac{2M}{r(p)} \right]^{1/2}, \quad \chi' \equiv (-\xi^a \xi_a)^{1/2}|_{p'} = \left[1 - \frac{2M}{r(p')} \right]^{1/2}. \quad (9.4.56)$$

However, the redshift corresponding to $\Delta\tau' > \Delta\tau$ we mentioned before refers to $(\lambda' - \lambda)/\lambda$. Based on (9.4.55), to find λ'/λ one only has to find $\tilde{\lambda}/\lambda$. In this case, only p' is involved, and we can deal with this by using the same approach of special relativity (see Sects. 7.2 and 7.5). This is essentially a problem of Doppler frequency shift, and so we can use (6.6.66a) directly, in which the γ can be derived as follows:

$$\gamma \equiv -g_{ab} Z^a \tilde{Z}^b = -g_{ab} (\partial/\partial\tau)^a \chi^{-1} (\partial/\partial t)^b = \chi^{-1} E.$$

(E is the energy of the timelike geodesic with Z^a as the tangent vector.) Then, from $\gamma = (1 - u^2)^{-1/2}$ we find the 3-speed of Z^a relative to \tilde{Z}^a , $u = \sqrt{E^2 - \chi^2}/E$. Plugging this into (6.6.66a) yields

$$\frac{\tilde{\lambda}}{\lambda} = \frac{E + \sqrt{E^2 - \chi^2}}{\chi}. \quad (9.4.57)$$

The above equation indicates that when p is infinitesimally approaching the event horizon, there exists an infinite (Doppler) redshift between the wavelengths measured by the observers Z^a and \tilde{Z}^a . Combining (9.4.55) and (9.4.57) we can find λ'/λ :

$$\frac{\lambda'}{\lambda} = \frac{\chi'(E + \sqrt{E^2 - \chi^2})}{\chi^2}. \quad (9.4.58)$$

The above equation can be viewed as a combination of the Doppler redshift and the gravitational redshift. By means of this equation we can also give a proof to a conclusion we claimed before—for Fig. 9.21 we have $\Delta\tau' > \Delta\tau$, and $\Delta\tau'/\Delta\tau$ increases as τ increases. Since $\Delta\tau$ and $\Delta\tau'$ can be interpreted as the periods of the light wave when it is emitted and received, we have

$$\frac{\Delta\tau'}{\Delta\tau} = \frac{\chi'(E + \sqrt{E^2 - \chi^2})}{\chi^2} > \frac{\chi'}{\chi^2} E > E. \quad (9.4.59)$$

The E in the above equation is the energy of the world line (geodesic) of a point on the surface of the collapsing star, i.e.,

$$E = -g_{ab}(\partial/\partial t)^a(\partial/\partial\tau)^b = -\chi g_{ab}\tilde{Z}^a(\partial/\partial\tau)^b = \chi\gamma,$$

where $\gamma \equiv -g_{ab}\tilde{Z}^a(\partial/\partial\tau)^b$. Extend this geodesic backwards to $r = \infty$, then $\chi = 1$ while γ is still greater than (or equal to) 1, and hence $E \geq 1$. From (9.4.59) we see that $\Delta\tau' > \Delta\tau$, and $\Delta\tau'/\Delta\tau$ increases when χ decreases, and thus $\Delta\tau'/\Delta\tau$ increases as τ increases. That is, as the star collapses, the light emitted from its surface will have a stronger and stronger redshift when it reaches an exterior observer.

[The End of Optional Reading 9.4.2]

Exercises

- ~9.1. Consider Taub's plane symmetric static spacetime, whose line element is (8.6.1'). By means of the Killing vector fields, write down the decoupled equations satisfied by the parametrization $t(\tau)$, $x(\tau)$, $y(\tau)$ and $z(\tau)$ of the timelike geodesic $\gamma(\tau)$ (the reader may refer to Sect. 9.1).
- 9.2. In Newton's theory of gravity, derive (9.3.18) directly using Fig. 9.8.
- ~9.3. Show that the OV equation of hydrostatic equilibrium (9.3.17) can be rewritten as

$$\left[1 - \frac{2m(r)}{r}\right]^{1/2} \frac{dp}{dr} = -(\rho + p)g, \quad (9.4.60)$$

where g represents the magnitude of the 4-acceleration $U^b\nabla_b U^a$ of a fluid particle.

Remark. Under the Newtonian approximation $[1 - 2m(r)/r]^{1/2} \cong 1$, $p \cong 0$, and (9.4.60) becomes $dp/dr \cong -\rho g$. Also, $g \cong m(r)/r^2$, and hence we get (9.3.18), i.e., $dp/dr \cong -\rho m(r)/r^2$.

- ~9.4. Show that (9.3.26) approximately goes back to (9.3.23) of Newton's theory of gravity when $R \gg M$.
- ~9.5. Find the relation between the Rindler coordinates t, x and the Lorentzian coordinates T, X in Minkowski spacetime.
- ~9.6. Which Killing vector field in Minkowski spacetime is the timelike Killing vector field $(\partial/\partial t)^a$ in Rindler spacetime?
- ~9.7. Find the magnitude $A \equiv (A^a A_a)^{1/2}$ of the 4-acceleration of a static observer in Schwarzschild spacetime. Hint: one may use the conclusion of Exercise 8.3, i.e., $A_a = \nabla_a \ln \chi$.
- ~9.8. Name the null geodesic represented by N_1 (or N_2) in Fig. 9.13a as N_1 (or N_2) for short. Show that: (1) the coordinate V (or U) is an affine parameter of the null geodesic N_1 (or N_2); (2) the coordinate r is an affine parameter of a radial null geodesic other than N_1 and N_2 .
- ~9.9. By introducing coordinates similar to the Kruskal coordinates, eliminate the coordinate singularity of the following line element at $r = R$:

$$ds^2 = -(1 - r^2/R^2)dt^2 + (1 - r^2/R^2)^{-1}dr^2 + r^2(d\theta^2 + \sin^2 \theta d\varphi^2), \quad R = \text{constant}.$$

- 9.10. Show that the maximally extended Schwarzschild spacetime has an s.p. curvature singularity. Hint: use (8.3.21).
- 9.11. Show that the N_1 in Fig. 9.13 is a null hypersurface. Hint: one only has to show that its normal vector n^a is null. Note that the equation of N_1 is $U = 0$, and hence its normal covector is $n_a \equiv \nabla_a U$.
- 9.12. Derive (9.4.51) from (9.4.50), and then derive (9.4.54).
- ~9.13. Write down the expression for the line element of the Schwarzschild metric in the outgoing Eddington-Finkelstein coordinate system $\{u, r, \theta, \varphi\}$ ($u \equiv t - r_*$).
- *9.14. Show that the ξ^a defined in terms of $(\partial/\partial V)^a$ and $(\partial/\partial U)^a$ [see (9.4.40)] is a null Killing vector field on N_1 and N_2 .
- *9.15. Transform Figs. 9.21, 9.22 and 9.23. Give another derivation for (9.4.58) by calculating the $\Delta\tau'/\Delta\tau$ in the figure. Hints: (1) $U \equiv -e^{(r_*-t)/4M}$ is a constant on each outgoing null geodesic. Along the world line of an exterior static observer and the world line of a freely falling observer on the star's surface, derive two expressions for the same dU (with $d\tau'$ and $d\tau$ respectively). Then putting an equal sign between the two expressions yields (9.4.58). (2) When writing the expression of dU in terms of $d\tau$ one needs to use the formulae of $dt/d\tau$ and $dr/d\tau$ expressed by the energy E , which can be obtained using the approach in Sect. 9.1.

Fig. 9.22 The relation between the wavelengths measured by Z'^a and \tilde{Z}^a is gravitational redshift; the relation between the wavelengths measured by Z^a and \tilde{Z}^a is Doppler redshift

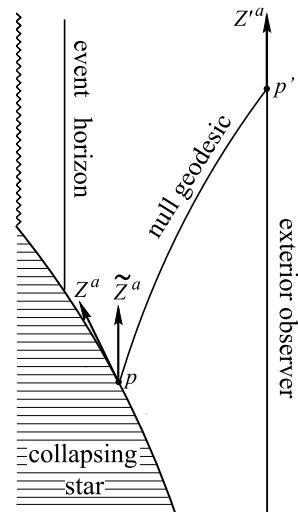
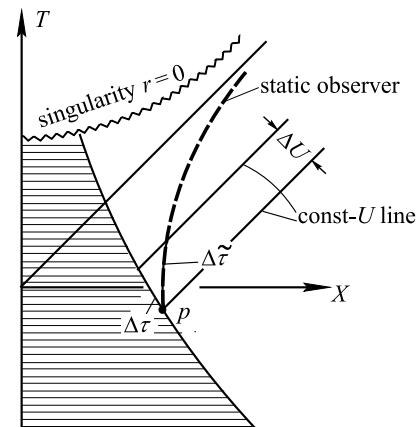


Fig. 9.23 Another method of deriving (9.4.58) (see Exercise 9.15)



References

- Chandrasekhar, S. (1939), *An Introduction to the Study of Stellar Structure*, University of Chicago Press, Chicago.
- Geroch, R. P. (1968), ‘What is a spacetime singularity in general relativity’, *Ann. Phys.* **48**, 526–540.
- Hawking, S. W. and Ellis, G. F. R. (1973), *The Large Scale Structure of Space-Time*, Cambridge University Press, Cambridge.
- Kruskal, M. D. (1960), ‘Maximal extension of Schwarzschild metric’, *Phys. Rev.* **119**, 1743–1745.
- Misner, C., Thorne, K. and Wheeler, J. (1973), *Gravitation*, W H Freeman and Company, San Francisco.
- Ni, W.-T. (2005), ‘Empirical foundations of relativistic gravity’, *Int. J. Mod. Phys. D* **14**, 901–922.
[arXiv:gr-qc/0504116](https://arxiv.org/abs/gr-qc/0504116).

- Ni, W.-T. (2016), ‘Solar-system tests of the relativistic gravity’, *Int. J. Mod. Phys. D* **25**(14), 1630003. [arXiv:1611.06025](https://arxiv.org/abs/1611.06025).
- Sachs, R. K. and Wu, H. (1977), *General Relativity for Mathematicians*, Springer-Verlag, New York.
- Shapiro, S. S., Davis, J. L., Lebach, D. E. and Gregory, J. S. (2004), ‘Measurement of the solar gravitational deflection of radio waves using geodetic very-long-baseline interferometry data, 1979–1999’, *Phys. Rev. Lett.* **92**, 121101.
- Wald, R. M. (1984), *General Relativity*, The University of Chicago Press, Chicago.
- Weinberg, S. (1972), *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*, John Wiley and Sons, New York.
- Will, C. M. (2018), *Theory and Experiment in Gravitational Physics*, Cambridge University Press, Cambridge.

Chapter 10

Cosmology I



Thoughts of the universe began ever since the dawn of mankind, full of mystery, imagination, and wisdom. Almost every sage has thought over, talked about, and drawn conclusions concerning the universe. However, it is only after the development of general relativity that cosmology became a genuine science. From the point of view of general relativity, the universe is the maximal spacetime containing everything in Nature, with its curvature on large scales and a distribution of matter satisfying the Einstein field equation.

Among the various branches of physics, cosmology is the most special one in the following sense: the object it concerns is unique—our universe. There are no other objects that could be compared with the universe. It is impossible to do experiments again and again as is done in other branches of physics, because the evolution of the universe cannot be replayed. The only way to study cosmology is to accumulate data from observations, to develop cosmological models in order to interpret these data, to speculate on the unknown past history of the universe, and to predict its future. There are many cosmological models. In this chapter only the mostly accepted one is introduced, known as the **standard cosmological model**¹ for its notable success.

There are still various problems in the standard model. Hence, it has been continuously amended ever since its birth. For example, an important amendment to it is inserting an “inflation” period in the very early universe. Furthermore, observations in 1998 showed that the universe is currently undergoing an *accelerating* expansion, which consequently also requires that the standard model must be amended. A new standard cosmological model is in development, although there are still open questions. We will introduce the inflationary model and the new standard cosmological model in Volume II.

¹ We will refer to the standard cosmological model as the “standard model” for short when there is no confusion with the Standard Model of particle physics.

Cosmology is an actively progressing subject. Therefore, even though the conclusions and data are being updated at the time of writing this book, they are still possibly already old fashioned when the book is published. For the latest progress of cosmology, readers have to refer to the new literature.

10.1 Kinematics of the Universe

10.1.1 *Cosmological Principle*

A fundamental postulate in the standard cosmological model, as well as other models, is the **cosmological principle**: at each moment, the cosmic space is homogeneous and isotropic when viewed on a very large scale.

Homogeneity of the cosmic space means that the physical properties are the same at every point in the space, with no single point being more special than any other. This is, of course, incorrect on ordinary scales, for stars might be here and there in the universe, but not everywhere. In fact, matter in the universe tends to be accumulated: matter is accumulated into stars, stars are accumulated into **galaxies**, (approximately 10^6 to 10^{13} stars are contained in a galaxy. The Milky Way is a quite ordinary galaxy, consisting of several hundred billion stars.) and some galaxies are accumulated into **clusters of galaxies** of various sizes, or even **superclusters**, However, based on observations on scales of 10^{10} ly (where ly denotes light-year) or larger, it is believed that the universe is homogeneous on a very large scale (one greater than 3×10^8 ly, referred to as a cosmological scale, which is large enough to contain many clusters of galaxies), with the density being constant from point to point. Here the density refers to the average density over a cosmological scale, the density obtained by smoothing matter in a volume that is small enough when viewed on a cosmological scale. Smoothing is a common trick in physics. For example, matter is not continuously distributed on microscopic scales (mainly concentrated in atomic nuclei), but it is regarded as being continuously distributed when viewed on macroscopic scales, so that a macroscopic density can be defined. If the macroscopic density is uniform from point to point, in a certain volume, then we say that the matter is homogeneously distributed in this volume. Here, a “point” in the volume is a “macroscopic point”, which is a volume that is small enough on macroscopic scales and large enough on microscopic scales, containing a large number of molecules.

By isotropy, we mean that there is a reference frame such that, for any observer in it, all directions appear the same, with no single direction being more special than others. This is also not true on ordinary scales, since we may see a star in one direction, but no star in another direction. However, it is accepted that there exists such a reference frame, as described above, if the universe is smoothed on a cosmological scale.

The postulate that cosmic space is homogeneous and isotropic at a large scale was suggested by Albert Einstein in 1917, when he applied general relativity to

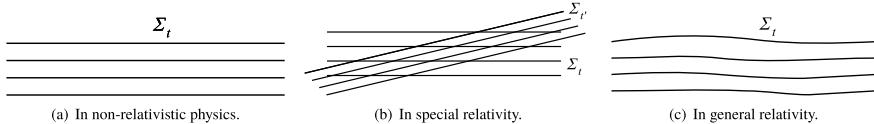


Fig. 10.1 A foliation of the spacetime, each slice represents the space at a certain time

cosmology. At that time, there was little observational data to support it. He suggested such a postulate in order to simplify the discussion. [Mach's principle also contributed to the development of this postulate, see Peebles (1993) pp. 10–16.]

The cosmological principle concerns the properties of the cosmic space. In pre-relativity physics, the word “space” is rather simple to understand. In relativity, however, it is not that simple. The key point is that spacetime is an absolute object in relativity, while space and time are related to a decomposition by an observer. For the same spacetime, different space and time are the result of different $3+1$ decompositions. To provide a precise interpretation of “spatial homogeneity”, one must first have a clear definition of “space”.

In pre-relativity physics, each surface Σ of absolute simultaneity is the “whole space” at a certain moment, see Fig. 6.10. In this sense the concept of space (as an absolute concept) is rather simple. In special relativity, given an inertial reference frame $\{t, x, y, z\}$, a constant- t surface Σ_t is the whole space at the time t relative to this inertial reference frame. In the above cases, each spacetime is foliated by constant- t surfaces. (This means that, for each spacetime point p , there is exactly one constant- t surface Σ_t such that $p \in \Sigma_t$.) In the former case, the foliation (slicing) is absolute (i.e., the foliation is unique, as shown in Fig. 10.1a), while in the latter, the foliation is relative (i.e. the foliations for different reference frames are different, as shown in Fig. 10.1b).

In special relativity, the foliation relative to a given inertial reference frame has the following properties:

① Each leaf (slice) is a connected spacelike hypersurface.

② The set $\{\Sigma_t\}$ of all leaves of the foliation is a 1-parameter family. That is, each real number t corresponds to a unique leaf Σ_t in the set, whose t is the coordinate time of this leaf relative to the given inertial reference frame. Thus a leaf is also called a specific “time”.

In general relativity, there are no global inertial reference frames in a curved spacetime. Instead, any foliation similar to the above is acceptable, with each leaf of the foliation identified with a time. To be more precise, in cosmology, a **foliation** (or **slicing**) $\{\Sigma_t\}$ of a spacetime (M, g_{ab}) is characterized by a smooth function τ on M satisfying the following conditions:

① $(d\tau)_a$ is a timelike 1-form with $(d\tau)_a Z^a > 0$ for any future-directed timelike vector field Z^a . It follows that each constant- τ surface is a spacelike hypersurface.

② For any real number t , the corresponding constant- τ surface, denoted by $\Sigma_t \equiv \{p \in M \mid \tau(p) = t\}$, is either empty or connected.

③ For each pair of real numbers t and t' , if both Σ_t and $\Sigma_{t'}$ are not empty, they must be diffeomorphic to each other.

It is easy to see that, for each $p \in M$, there exists a unique Σ_t such that $p \in \Sigma_t$. In fact, $p \in \Sigma_{\tau(p)}$. Each surface Σ_t is called a “time”, which is a **leaf** (or **slice**) of the foliation, see Fig. 10.1c.

To be specific, the foliation $\{\Sigma_t\}$ as in the above is also called the **foliation associated with τ** . Let $\{\Sigma'_t\}$ be a foliation of M associated with a function τ' . Then, since both τ and τ' are smooth maps from the connected manifold M to \mathbb{R} , and since both $(d\tau)_a$ and $(d\tau')_a$ are nonvanishing on M , the images $I \equiv \tau[M]$ and $I' \equiv \tau'[M]$ are both open intervals. In this chapter, we regard $\{\Sigma_t\}$ and $\{\Sigma'_t\}$ as the same (or an equivalent) foliation if there is a diffeomorphism $f : I \rightarrow I'$ such that $\tau' = f \circ \tau$. In other words, the foliations $\{\Sigma_t\}$ and $\{\Sigma'_t\}$ are equivalent if they differ from each other by a reparametrization. Later, when we say a spacetime admits a unique homogeneous foliation, it will be in this sense.

For a more detailed discussion on the topics of foliation as well as 3 + 1 decomposition, the reader may refer to Sect. 14.4 in Volume II.

In relativity, acceptable foliations for a spacetime satisfying the above conditions are not unique. This is how the concepts of space and time in relativity have arbitrariness. If the spacetime has some nontrivial symmetries, foliations adapted to these symmetries are more convenient. In fact, as a special case with zero curvature, Minkowski spacetime admits various foliations, among which a foliation consisting of surfaces of simultaneity relative to an inertial reference frame is the most accepted, because it is associated with the symmetries of Minkowski spacetime.

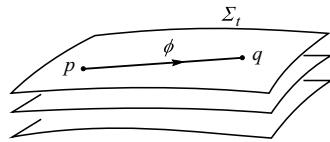
Spatial homogeneity and spatial isotropy of a spacetime are closely related to some intrinsic symmetries of the spacetime. By spatial homogeneity, we refer to the existence of a foliation of the spacetime such that, at all points in the same arbitrary leaf of the foliation, the geometric properties and physical properties are the same. Hence each leaf in this foliation is called a surface of homogeneity. Such a foliation is adapted to the intrinsic symmetries. Although other foliations are acceptable, they are not convenient for use in that some of their leaves are not surfaces of homogeneity. Unless otherwise stated, when we talk about spaces in cosmology they are all surfaces of homogeneity. Spatial isotropy is relative to a reference frame. If there exists a reference frame in spacetime where no observer can find any spatial direction (a direction orthogonal to the world line) distinct from other spatial directions in a local experiment at a certain time, then we say that the spacetime is spatially isotropic.

10.1.2 Spacial Geometries of the Universe

In this subsection we will show that on each surface of homogeneity, there are only three kinds of possible geometry satisfying the cosmological principle. This conclusion will significantly simplify the successive discussions.

The cosmological principle assumes that the universe is spatially homogeneous and spatially isotropic in both physics and geometry. To be precise, and to be

Fig. 10.2 Figure for defining spatial homogeneity



convenient for discussing the spatial geometry of the universe, we shall first define some geometric concepts, including spatial homogeneity and spatial isotropy, using mathematical language.

Definition 1 A generalized Riemannian space (M, g_{ab}) is said to be **homogeneous** if, for any $p, q \in M$, there exists an isometry $\phi : M \rightarrow M$ of g_{ab} such that $\phi(p) = q$.

An embedding submanifold $i : S \rightarrow M$ is said to be **homogeneous** if (S, h_{ab}) is a homogeneous generalized Riemannian space, where $h_{ab} = i^*g_{ab}$ is the induced metric. For convenience, we will also refer to the image $i[S]$ as a homogeneous submanifold of M .

A foliation $\{\Sigma_t\}$ of a spacetime (M, g_{ab}) is called a **homogeneous foliation** if each leaf in it is a homogeneous submanifold.

A spacetime (M, g_{ab}) is said to be **spatially homogeneous** if it admits a homogeneous foliation $\{\Sigma_t\}$ (see Fig. 10.2). Each Σ_t in the foliation is called a **surface of homogeneity**.

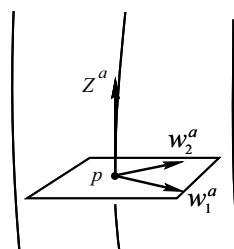
Definition 2 A reference frame \mathcal{R} in a spacetime (M, g_{ab}) is said to be **spatially isotropic**, or **isotropic** for short, if for any point on the world line of any observer (with a 4-velocity Z^a) in \mathcal{R} , and for any two spatial vectors w_1^a and w_2^a at p with the same magnitude, there exists an isometry $\psi : M \rightarrow M$ of g_{ab} , such that $\psi(p) = p$, $\psi_*Z^a = Z^a$ and $\psi_*w_1^a = w_2^a$ (see Fig. 10.3).

An observer in an isotropic reference frame is called an **isotropic observer**.

A spacetime admitting an isotropic reference frame is called an **isotropic spacetime**.

For the comparison of scales, a galaxy to the universe is as a drop to the ocean. Thus, a galaxy is treated as a world line in the spacetime. One may take the conjecture that every galaxy is an isotropic observer (observations indicate that this conjecture is almost true, with only a tiny deviation). In the following, galaxies will refer to isotropic observers unless stated otherwise.

Fig. 10.3 Figure for defining an isotropic observer



For a spacetime that is both spatially homogeneous and isotropic, it is natural to ask about the relation between \mathcal{R} , its isotropic reference frame, and Σ_t , a surface of homogeneity. We certainly hope that the world lines of the observers in \mathcal{R} are orthogonal to Σ_t . However, this is not necessarily true for Minkowski spacetime. This is because any inertial reference frame in Minkowski spacetime is isotropic, and the family of surfaces of simultaneity relative to an inertial reference frame is a homogeneous foliation, but the world line of an observer stationary in one inertial reference frame is not orthogonal to the surface of simultaneity relative to another inertial reference frame. However, if a spacetime that is both spatially homogeneous and isotropic possesses a unique family of surfaces of homogeneity, then the orthogonality holds. See Proposition 10.1.1, which follows.

Proposition 10.1.1 *If an isotropic spacetime (M, g_{ab}) has a unique homogeneous foliation $\{\Sigma_t\}$, then the world line of an isotropic observer is orthogonal to each leaf in the homogeneous foliation.*

[Optional Reading 10.1.1]

Before giving the proof of Proposition 10.1.1, we first prove Lemma 10.1.2 and Proposition 10.1.3.

Lemma 10.1.2 *Suppose $\Sigma \subset M$ is a homogeneous submanifold of the generalized Riemannian space (M, g_{ab}) . Let $\psi : M \rightarrow M$ be an arbitrary isometry of g_{ab} , then $\psi[\Sigma]$ is also homogeneous.*

Proof For clarity, we will distinguish an embedded submanifold (as a map) and its image in this proof. Consider a generalized Riemannian space (S, h_{ab}) so that the map $i : S \rightarrow M$ is the embedded submanifold, whose image $i[S] = \Sigma$. Then $h_{ab} = i^*g_{ab}$. Since ψ is an isometry, according to Theorem 4.4.5, $i' = \psi \circ i : S \rightarrow M$ is also an embedded submanifold of M , and the corresponding induced metric $h'_{ab} = i'^*g_{ab}$ is equal to h_{ab} .

The image Σ of $i : S \rightarrow M$ being homogeneous means that (S, h_{ab}) is a homogeneous generalized Riemannian space. Now that $h'_{ab} = h_{ab}$, $i' : S \rightarrow M$ is also a homogeneous submanifold. It is easy to see that $\psi[\Sigma]$ is the image of $i' : S \rightarrow M$. Therefore, $\psi[\Sigma]$ is also homogeneous. \square

Proposition 10.1.3 *Suppose $\{\Sigma_t\}$ is a homogeneous foliation of a spacetime (M, g_{ab}) associated with a function τ . Let $\psi : M \rightarrow M$ be an isometry of (M, g_{ab}) . If ψ preserves the future direction (time orientation), then $\{\psi[\Sigma_t]\}$ is a homogeneous foliation of (M, g_{ab}) associated with the function $(\psi^{-1})^*\tau$; otherwise, $\{\psi[\Sigma_t]\}$ is a homogeneous foliation of (M, g_{ab}) associated with the function $-(\psi^{-1})^*\tau$.*

Proof First, we show that $\{\psi[\Sigma_t]\}$ satisfies the three conditions of a foliation. Let $\tau' = (\psi^{-1})^*\tau$, then τ' is a smooth function on M . Since ψ is an isometry, $(d\tau')_a = (\psi^{-1})^*(d\tau)_a$ is nonvanishing on M . For an arbitrary future-directed timelike vector field Z^a on M and any $p \in M$,

$$[(d\tau')_a Z^a]_p = [(\psi^{-1})^*(d\tau)_a](Z^a|_p) = (d\tau)_a[(\psi^{-1})_*(Z^a|_p)].$$

If the isometry ψ preserves the future direction, so does ψ^{-1} . Thus, $(\psi^{-1})_*(Z^a|_p)$ is a future-directed timelike vector at $\psi^{-1}(p)$. It follows that $(d\tau)_a(\psi^{-1})_*(Z^a|_p) > 0$, i.e., $[(d\tau')_a Z^a]_p > 0$.

For any $p \in M$, $\tau'|_p = (\tau \circ \psi^{-1})|_p = \tau|_{\psi^{-1}(p)}$. Therefore, $\forall t \in \mathbb{R}$, the necessary and sufficient condition of $p \in \psi[\Sigma_t]$ is $\psi^{-1}(p) \in \Sigma_t$. The latter is equivalent to $t = \tau|_{\psi^{-1}(p)} = \tau'|_p$. Hence, $p \in \psi[\Sigma_t]$ if and only if p is in the constant- τ' surface with $\tau' = t$. That is, $\forall t \in \mathbb{R}$, $\psi[\Sigma_t]$ is a constant- τ' surface. Since $\{\Sigma_t\}$ is a foliation, each leaf is either empty or connected. And since ψ is a diffeomorphism, we see that $\psi[\Sigma_t]$ is either empty or connected.

$\{\Sigma_t\}$ being a foliation also means that any two leafs Σ_t and $\Sigma_{t'}$ are diffeomorphic to each other. Since ψ is a diffeomorphism, it is easy to see that $\psi[\Sigma_t]$ and $\psi[\Sigma_{t'}]$ are also diffeomorphic to each other. Therefore, $\{\psi[\Sigma_t]\}$ is a foliation of (M, g_{ab}) associated with the function $\tau' = (\psi^{-1})^*\tau$.

Since $\{\Sigma_t\}$ is a homogeneous foliation, each $\Sigma_t \in \{\Sigma_t\}$ is a homogeneous hypersurface of M . Then, according to Lemma 10.1.2, $\psi[\Sigma_t]$ is homogeneous. As a consequence, $\{\psi[\Sigma_t]\}$ is a homogeneous foliation.

In a similar manner, one can show that if ψ does not preserve the future direction, then $\{\psi[\Sigma_t]\}$ is a homogeneous foliation associated with the function $-(\psi^{-1})^*\tau$. \square

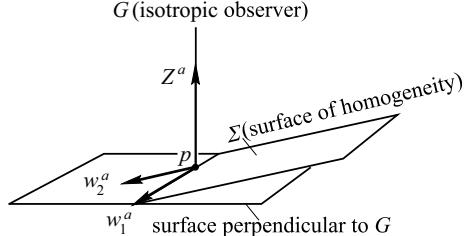
Proof of Proposition 10.1.1 Suppose p is a point on the world line of an isotropic observer G , and Σ is the surface of homogeneity containing p which is not orthogonal to the 4-velocity Z^a of G at p (see Fig. 10.4). Suppose V_p is the 4-dimensional tangent space at p , and W_p is the linear subspace of V_p orthogonal to Z^a . Then the elements in W_p are spatial vectors (with respect to G) at p . Let $w_1^a \in W_p$ be a unit vector tangent to Σ . Since Σ is not orthogonal to Z^a , there also exists a unit vector $w_2^a \in W_p$ that is not tangent to Σ .

Let ψ be an arbitrary isometry of (M, g_{ab}) such that $\psi(p) = p$, and $\psi_* Z^a = Z^a$. Then, according to Proposition 10.1.3, $\{\Sigma_t\}$ being a homogeneous foliation assures that $\{\psi[\Sigma_t]\}$ is also a homogeneous foliation. Due to the uniqueness of homogeneous foliations, we have $\{\psi[\Sigma_t]\} = \{\Sigma_t\}$. Especially, for the leaf Σ containing p , $\psi[\Sigma]$ is also a leaf in the same foliation, which contains $\psi(p) = p$. Hence, $\psi[\Sigma] = \Sigma$. Since w_1^a is tangent to Σ , $\psi_* w_1^a$ is tangent to $\psi[\Sigma] = \Sigma$, and thus $\psi_* w_1^a \neq w_2^a$. Consequently, there is no isometry ψ such that $\psi(p) = p$ and $\psi_* w_1^a = w_2^a$, which contradicts the fact that G is an isotropic observer. \square

[The End of Optional Reading 10.1.1]

Corollary 10.1.4 *If an isotropic spacetime has a unique homogeneous foliation, then*

Fig. 10.4 Figure for the proof of Proposition 10.1.1



- (1) its isotropic reference frame is unique;
- (2) an isometry of the spacetime maps one isotropic observer to another.

In addition to the cosmological principle, it is also assumed that the homogeneous foliation of the cosmic spacetime (i.e., the universe) is unique. Therefore, the world lines of isotropic observers are orthogonal to the surfaces of homogeneity, and the unique isotropic reference frame is called the **cosmic rest frame** (it corresponds to the **comoving reference frame** in Sect. 6.5). In this way there is a unique “orthogonal 3 + 1 decomposition” of the cosmic spacetime, in which each surface of homogeneity is the whole space at a time, and each world line of an isotropic observer (which is orthogonal to the surfaces of homogeneity) represents the whole history of a spatial point. Now we will discuss the 3-dimensional geometry of the space (surface of homogeneity) at an arbitrary time.

Proposition 10.1.5 Suppose h_{ab} is the metric on a surface Σ_t of homogeneity induced by the metric g_{ab} of the cosmic spacetime. Let \hat{R}_{abc}^d be the curvature tensor of h_{ab} , and $\hat{R}_{abcd} \equiv h_{de}\hat{R}_{abc}^e$, then there exists a constant K such that

$$\hat{R}_{abcd} = 2Kh_{c[a}h_{b]d}. \quad (10.1.1)$$

Proof Let $\Lambda_p(2)$ be the collection of all the 2-forms on Σ_t at an arbitrary $p \in \Sigma_t$, where Σ_t is treated as an independent 3-dimensional manifold. According to Theorem 5.1.3, $\Lambda_p(2)$ is a 3-dimensional vector space. Let $\hat{R}_{ab}^{cd} \equiv h^{ce}\hat{R}_{abe}^d$, then, $\forall Y_{cd} \in \Lambda_p(2)$, we have $\hat{R}_{ab}^{cd}Y_{cd} \in \Lambda_p(2)$. Hence, \hat{R}_{ab}^{cd} is a linear map from $\Lambda_p(2)$ to itself, namely a linear operator on $\Lambda_p(2)$. It follows from the symmetry of the curvature tensor, $\hat{R}_{abcd} = \hat{R}_{cdab}$, that \hat{R}_{ab}^{cd} can be regarded as a symmetric operator² L on $\Lambda_p(2)$, i.e., the corresponding matrix of L in an orthonormal basis of $\Lambda_p(2)$ is symmetric (for a proof, see Optional Reading 10.1.2). Hence, L is diagonalizable, i.e., one can choose a basis of $\Lambda_p(2)$ such that each vector in this basis is an eigenvector of L . Due to the isotropy and the uniqueness of the homogeneous foliation, it can be proved that all the eigenvalues of L are equal (see Optional Reading 10.1.2). Thus, L could only be the scalar product of a real number $2K$ and the identity map I on $\Lambda_p(2)$. That is,

$$L = 2KI, \quad K \in \mathbb{R}. \quad (10.1.2)$$

Also, it follows from

$$\delta_a^{[c}\delta_b^{d]}Y_{cd} = \delta_a^c\delta_b^dY_{[cd]} = \delta_a^c\delta_b^dY_{cd} = Y_{ab}$$

that the identity map I on $\Lambda_p(2)$ corresponds to the tensor $\delta_a^{[c}\delta_b^{d]}$. Thus, (10.1.2) can be rewritten as a equality of tensors

² In linear algebra, a linear operator L on a real vector space V is said to be **symmetric** or **self-adjoint** if $(u, Lv) = (Lu, v)$ for arbitrary $u, v \in V$. More discussion on the self-adjointness of linear operators will be given in Appendix B in Volume II.

$$\hat{R}_{ab}^{cd} = 2K\delta_a^{[c}\delta_b^{d]} . \quad (10.1.3)$$

So far the above equation is valid at a point $p \in \Sigma_t$. Since p is arbitrary, (10.1.3) holds pointwise on Σ_t with K being a scalar field. On account of spatial homogeneity, K is required to be a constant. Then, contracting both sides of (10.1.3) with $h_{ce}h_{df}$ and noticing that $h_{ce}h_{df}\hat{R}_{ab}^{cd} = \hat{R}_{abef}$, we obtain (10.1.1). \square

[Optional Reading 10.1.2]

Now we prove the following two conclusions we used in the proof above:

(1) For each $p \in \Sigma_t$, when \hat{R}_{ab}^{cd} is regarded as a linear operator on $\Lambda_p(2)$, it is a symmetric operator L .

(2) $L = 2KI$ with $K \in \mathbb{R}$, where I is the identity map on $\Lambda_p(2)$.

Before talking about $\Lambda(2)$, let us first consider an n -dimensional vector space V over \mathbb{R} with an inner product (\cdot, \cdot) . According to the theory of linear algebra, for any symmetric operator L on V , there exists an orthonormal basis of V formed by the eigenvectors of L . $\Lambda_p(2)$ is a 3-dimensional vector space over \mathbb{R} . In order to apply the above conclusion to $\Lambda_p(2)$ and its linear operator \hat{R}_{ab}^{cd} , an inner product must be defined on $\Lambda_p(2)$: for any $X_{ab}, Y_{ab} \in \Lambda_p(2)$, the inner product is defined by $(X, Y) := X^{ab}Y_{ab}$, where $X^{ab} = h^{ac}h^{bd}X_{cd}$. It is obvious that the inner product (\cdot, \cdot) of $\Lambda_p(2)$ is symmetric and bilinear. Since h_{ab} is positive definite, (\cdot, \cdot) is positive definite. Thus, (\cdot, \cdot) is indeed an inner product.

For each $X_{ab} \in \Lambda_p(2)$, $\hat{R}_{ab}^{cd}X_{cd}$ will be denoted by LX , with abstract indices omitted. Then, for arbitrary $X_{ab}, Y_{ab} \in \Lambda_p(2)$,

$$(X, LY) = X^{ab}(LY)_{ab} = X^{ab}\hat{R}_{ab}^{cd}Y_{cd} = \hat{R}_{abcd}X^{ab}Y^{cd},$$

$$(LX, Y) = (LX)^{ab}Y_{ab} = \hat{R}^{abcd}X_{cd}Y_{ab} = \hat{R}_{abcd}X^{cd}Y^{ab} = \hat{R}_{cdab}X^{ab}Y^{cd}.$$

It follows from $\hat{R}_{abcd} = \hat{R}_{cdab}$ that $(X, LY) = (LX, Y)$. This indicates that the linear operator L (i.e., \hat{R}_{ab}^{cd}) is a symmetric operator on $\Lambda_p(2)$. Therefore, there exists an orthonormal basis of $\Lambda_p(2)$ formed by the eigenvectors of L . In other words, the corresponding matrix of L in this basis is diagonal, and each diagonal element is the eigenvalue of a corresponding eigenvector.

Let I be the identity map on $\Lambda_p(2)$. In order to show that $L = 2KI$ with $K \in \mathbb{R}$, we only have to show that all the eigenvalues of L are equal and then denote them by $2K$, see the following proposition:

Proposition 10.1.6 Suppose $Y_{ab}, Y'_{ab} \in \Lambda_p(2)$ are two arbitrary eigenvectors of \hat{R}_{ab}^{cd} , and λ and λ' are the corresponding eigenvalues, i.e.,

$$\hat{R}_{ab}^{cd}Y_{cd} = \lambda Y_{ab}, \quad \hat{R}_{ab}^{cd}Y'_{cd} = \lambda' Y'_{ab}. \quad (10.1.4)$$

Then, $\lambda' = \lambda$ is assured by the isotropy and the uniqueness of the homogeneous foliation.

Proof Suppose (Σ_t, h_{ab}) is a 3-dimensional Riemannian space. Let W_p be the tangent space of Σ_t at $p \in \Sigma_t$. Then $(W_p, h_{ab}|_p)$ is a 3-dimensional vector space together with a positive definite metric, and $\Lambda_p(2)$ is nothing but the set of all the 2-forms on W_p . Denote the dual form of $Y_{ab} \in \Lambda_p(2)$ (which is a 1-form) by w_a . Then, according to (5.6.1), $w_c = Y^{ab}\hat{\varepsilon}_{abc}/2$, where $\hat{\varepsilon}_{abc}$ is the volume element associated with h_{ab} . Raising the index of w_c by h^{ac} , we obtain a spatial vector $w^a = \hat{\varepsilon}^{abc}Y_{bc}/2$. Similarly, there is also $w'^a = \hat{\varepsilon}^{abc}Y'_{bc}/2$. We assume that Y_{ab} and Y'_{ab} are chosen such that w^a and w'^a have the same magnitudes, then it follows from the fact that the cosmic spacetime is isotropic that there exists an isometry ψ of (M, g_{ab})

satisfying $\psi(p) = p$ and $\psi_* w^a = w'^a$. Since the homogeneous foliation is unique, we have $\psi[\Sigma_t] = \Sigma_t$ for the above isometry ψ . Hence, the restriction of ψ on Σ_t is an isometry $\psi|_{\Sigma_t} : \Sigma_t \rightarrow \Sigma_t$ of h_{ab} . For simplicity, we still denote $\psi|_{\Sigma_t}$ by ψ . Then, $\psi^* \hat{e}_{abc} = \hat{e}_{abc}$ and $\psi^* \hat{R}_{ab}^{cd} = \hat{R}_{ab}^{cd}$. It follows from the forms Y_{ab} and w_c being dual to each other that $Y_{ab} = \hat{e}_{cab} w^c$. Similarly, $Y'_{ab} = \hat{e}_{cab} w'^c$. Hence,

$$\psi^* Y'_{ab} = \psi^*(\hat{e}_{cab} w'^c) = \hat{e}_{cab} \psi_*^{-1} w'^c = \hat{e}_{cab} w^c = Y_{ab}. \quad (10.1.5)$$

On the other hand, from (10.1.4) we can see that

$$\psi^*(\hat{R}_{ab}^{cd} Y'_{cd}) = \psi^*(\lambda' Y'_{ab}). \quad (10.1.6)$$

Now let us look at both sides of (10.1.6):

$$\text{l.h.s. of (10.1.6)} = \hat{R}_{ab}^{cd} \psi^* Y'_{cd} = \hat{R}_{ab}^{cd} Y_{cd} = \lambda Y_{ab}, \quad (10.1.7)$$

where we used $\psi^* \hat{R}_{ab}^{cd} = \hat{R}_{ab}^{cd}$ in the first equality, (10.1.5) in the second equality, and (10.1.4) in the third equality. Also,

$$\text{r.h.s. of (10.1.6)} = \lambda' \psi^* Y'_{ab} = \lambda' Y_{ab}. \quad (10.1.8)$$

Combining (10.1.6), (10.1.7) and (10.1.8) yields $\lambda' = \lambda$. \square

[The End of Optional Reading 10.1.2]

[Optional Reading 10.1.3]

In the proof of Proposition 10.1.5, K being a constant on Σ_t follows from the spatial homogeneity. However, in order to show that K is a constant on Σ_t , spatial homogeneity is in fact not necessary, as shown in the following. Let ∇_a be the derivative operator on Σ_t associated with h_{ab} . Then, applying the Bianchi identity (3.4.8) to (10.1.2) yields $0 = \nabla_{[e} \hat{R}_{ab]cd} = 2h_{c[a} h_{b|d]} \nabla_{e]} K$. Contracting both sides with $h^{ad} h^{cb}$, we have $\nabla_e K = 0$ (where we have considered that the dimension n of Σ_t satisfies $n \geq 3$). This indicates that K is a constant on Σ_t , since Σ_t is connected as a leaf in a foliation. Therefore, when we only care about the geometry of Σ_t , spatial homogeneity is not required for obtaining the conclusion in Proposition 10.1.5.

[The End of Optional Reading 10.1.3]

A generalized Riemannian space (M, g_{ab}) is called a **space of constant curvature** if there exists a constant K such that the Riemann curvature tensor satisfies

$$R_{abcd} = 2K g_{c[a} g_{b]d}. \quad (10.1.1')$$

According to a proposition in Appendix J, ① a space of constant curvature is locally maximally symmetric; that is, the dimension of its isometry group (which may only be a local group), which is also the number of independent Killing vector fields, is $n(n - 1)/2$, where n is the spatial dimension. ② Two spaces of constant curvature with the same dimension, metric signature and K are (locally) isometric; that is, they have the same local geometry. Equation (10.1.1) indicates that each surface of homogeneity of the universe, (Σ_t, h_{ab}) , is a space of constant curvature.³ Therefore,

³ As shown in the proof of Proposition 10.1.6, the isotropy and the uniqueness of the homogeneous foliation is a sufficient condition for (Σ_t, h_{ab}) to be a space of constant curvature. In fact, the

if we can list out the h_{ab} corresponding to each real number K , we can exhaust all the possible (local) geometries of Σ_t .

Speaking of the geometries with maximal symmetry, the first thing one may have in mind is a flat metric. For a flat metric, its curvature tensor vanishes, which trivially satisfies (10.1.3) and (10.1.1') with $K = 0$. Hence, on a surface Σ_t of homogeneity, the line element with $K = 0$ can be expressed by means of Cartesian coordinates as

$$dl^2 = dx^2 + dy^2 + dz^2. \quad (10.1.9)$$

Of course, for $K \neq 0$ we have $\hat{R}_{ab}^{cd} \neq 0$, and thus it is impossible for h_{ab} to be flat when $K \neq 0$. Since the metric is maximally symmetric, besides the flat one, it is also natural to think about a spherically symmetric metric. Usually, a spherically symmetric metric refers to the metric on a 2-sphere, which is induced by the δ_{ab} of the 3-dimensional Euclidean space $(\mathbb{R}^3, \delta_{ab})$. The corresponding line element can be expressed as $d\theta^2 + \sin^2 \theta d\varphi^2$. But now we are looking for a spherically symmetric metric on a 3-dimensional space, which can be induced on a 3-sphere in 4-dimensional Euclidean space $(\mathbb{R}^4, \delta_{ab})$ by δ_{ab} . Let x, y, z and w be the Cartesian coordinates on \mathbb{R}^4 , then the equation of a 3-sphere (denoted by $S_{\bar{R}}$) reads

$$x^2 + y^2 + z^2 + w^2 = \bar{R}^2, \quad (10.1.10)$$

where $\bar{R} > 0$ is the radius of the 3-sphere. Similar to the 3-dimensional Euclidean space, the spherical coordinates in the 4-dimensional Euclidean space, R, ψ, θ and φ , are defined by

$$\begin{aligned} x &= R \sin \psi \sin \theta \cos \varphi, \\ y &= R \sin \psi \sin \theta \sin \varphi, \\ z &= R \sin \psi \cos \theta, \\ w &= R \cos \psi. \end{aligned} \quad (10.1.11)$$

Then, the line element of the 4-dimensional Euclidean space can be expressed as

$$ds^2 = dx^2 + dy^2 + dz^2 + dw^2 = dR^2 + R^2[d\psi^2 + \sin^2 \psi (d\theta^2 + \sin^2 \theta d\varphi^2)].$$

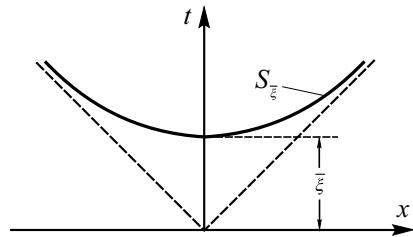
From (10.1.10) and (10.1.11) we can see that, on a 3-sphere $S_{\bar{R}}$ with radius \bar{R} , we have $R = \bar{R}$ and $dR = 0$. Thus, the line element on $S_{\bar{R}}$ induced by the line element of the 4-dimensional Euclidean space is

$$dl^2 = \bar{R}^2 [d\psi^2 + \sin^2 \psi (d\theta^2 + \sin^2 \theta d\varphi^2)]. \quad (10.1.12)$$

From the above line element, one can find that the curvature of $S_{\bar{R}}$ reads $\hat{R}_{ab}^{cd} = 2\bar{R}^{-2}\delta_a^{[c}\delta_b^{d]}$ (Exercise 10.1). Hence, the curvature of the 3-sphere $S_{\bar{R}}$ satisfies

uniqueness of the homogeneous foliation is not necessary, since Minkowski spacetime is obviously a space of constant curvature whose homogeneous foliation is not unique.

Fig. 10.5 A circular hyperboloid in Minkowski spacetime (with two dimensions suppressed)



(10.1.3) with $K = \bar{R}^{-2}$. Since \bar{R}^{-2} can be any positive number, the 3-spheres with various radii exhaust the local geometries of the 3-dimensional spaces of constant curvature with $K > 0$.

For spaces of constant curvature with $K < 0$, let us consider a 3-dimensional circular hyperboloid (denoted by $S_{\bar{\xi}}$, shown in Fig. 10.5 with two dimensions suppressed) in 4-dimensional Minkowski spacetime, determined by the following equation:

$$t^2 - x^2 - y^2 - z^2 = \bar{\xi}^2, \quad (10.1.13)$$

where t , x , y and z are the Lorentzian coordinates, and $\bar{\xi}$ is a positive constant. In the region of Minkowski spacetime where $t^2 - x^2 - y^2 - z^2 > 0$, we can define hyperbolic coordinates ξ , ψ , θ and φ by the following equations:

$$\begin{aligned} x &= \xi \sinh \psi \sin \theta \cos \varphi, \\ y &= \xi \sinh \psi \sin \theta \sin \varphi, \\ z &= \xi \sinh \psi \cos \theta, \\ t &= \xi \cosh \psi. \end{aligned} \quad (10.1.14)$$

Then, the 4-dimensional Minkowski line element can be expressed in the above region as

$$\begin{aligned} ds^2 &= -dt^2 + dx^2 + dy^2 + dz^2 \\ &= -d\xi^2 + \xi^2 [d\psi^2 + \sinh^2 \psi (d\theta^2 + \sin^2 \theta d\varphi^2)]. \end{aligned} \quad (10.1.15)$$

From (10.1.14) we can see that on the 3-dimensional hyperboloid $S_{\bar{\xi}}$ defined by (10.1.13), we have $\xi = \bar{\xi}$ and $d\xi = 0$, and hence the line element on $S_{\bar{\xi}}$ induced by the 4-dimensional Minkowski line element reads

$$dl^2 = \bar{\xi}^2 [d\psi^2 + \sinh^2 \psi (d\theta^2 + \sin^2 \theta d\varphi^2)]. \quad (10.1.16)$$

From the above line element, one can find that the the curvature of $S_{\bar{\xi}}$ is (left as an exercise)

$$\hat{R}_{ab}^{cd} = -2\bar{\xi}^{-2} \delta_a^{[c} \delta_b^{d]}. \quad (10.1.17)$$

Hence, the curvature on the 3-dimensional hyperboloid $S_{\bar{\xi}}$ satisfies (10.1.3) with $K = -\bar{\xi}^{-2}$. Since $\bar{\xi}^{-2}$ could be any positive number, the 3-dimensional hyperboloid $S_{\bar{\xi}}$ with various positive $\bar{\xi}$ exhausts the local geometries of the 3-dimensional spaces of constant curvature with $K < 0$.

Summary. As a consequence of the cosmological principle, at any time of the universe (i.e., for any surface of homogeneity), there are only three kinds of possible local spatial geometries, described by the following three kinds of metrics:

(a) 3-dimensional spherical metric, whose line element can be expressed in terms of the spherical coordinates ψ, θ and φ as

$$dl^2 = K^{-1} [d\psi^2 + \sin^2 \psi (d\theta^2 + \sin^2 \theta d\varphi^2)], \quad K > 0. \quad (10.1.18)$$

(b) 3-dimensional flat metric, whose line element can be expressed in terms of the Cartesian coordinates as

$$dl^2 = dx^2 + dy^2 + dz^2, \quad K = 0. \quad (10.1.19)$$

In terms of the spherical coordinates ψ, θ and φ , the above line element can also be written in a form similar to (10.1.18)

$$dl^2 = d\psi^2 + \psi^2 (d\theta^2 + \sin^2 \theta d\varphi^2), \quad (10.1.19')$$

(c) 3-dimensional hyperbolic metric, whose line element can be expressed in terms of the hyperbolic coordinates ψ, θ and φ as

$$dl^2 = -K^{-1} [d\psi^2 + \sinh^2 \psi (d\theta^2 + \sin^2 \theta d\varphi^2)], \quad K < 0. \quad (10.1.20)$$

Is the universe spatially finite? Throughout history, this question has been answered in both the affirmative and the negative. Each of these conceptions have dominated the mainstream in certain periods, with the amounts of time almost equal. Now, as it is certain that the spatial geometry of the universe can be classified into the above three cases, with some further requirements on the global topology (see the paragraph below) of the universe, this question becomes absolutely clear. In case (a), the space of the universe is a 3-dimensional sphere, resulting in a “closed universe” whose volume is finite. Although the universe is “finite” in this case, it is “boundless” since a sphere has no boundaries. In case (b) and (c), the space of the universe is respectively a 3-dimensional Euclidean space and a 3-dimensional hyperboloid, each resulting in an “open universe” with an infinite volume, and hence we say that the universe is “infinite” in these two cases. However, which case does our universe really belong to? We will discuss this problem in Sect. 10.3 and Chap. 15 in Volume II.

It is necessary to elucidate that a space of constant curvature only requires its metric to satisfy no more than condition (10.1.1') Especially, there is no condition for its global topological structure. Take a surface (Σ_t, h_{ab}) of homogeneity in the universe as an example. In the case of $K > 0$, (10.1.1) only leads to the conclusion

that the metric h_{ab} is a “3-dimensional spherical metric” satisfying (10.1.18). It does not indicate that Σ_t itself is a 3-sphere, since Σ_t may only be part of a 3-sphere, or, for example, a space obtained by identifying antipodal points of a 3-sphere (namely a quotient space of a 3-sphere), etc. All such spaces share the same local geometry. In the case of $K = 0$, it is possible that Σ_t is a space obtained by identifying a pair of points in \mathbb{R}^3 whenever the differences of their x -coordinates, y -coordinates and z -coordinates are all integers. For such a space Σ_t , which is known as a quotient space of \mathbb{R}^3 , its metric is still flat, and its volume is finite (being exactly 1). By virtue of the cosmological principle, Σ_t cannot be a proper subset of a 3-sphere, or of a 3-dimensional Euclidean space, or of a 3-dimensional hyperboloid, but the quotient spaces we mentioned above still satisfy the cosmological principle. Nevertheless, from the perspective of physics, the quotient spaces are not regarded as being natural. Then, this finally leads to the conclusion about the global spatial geometry of the universe: when $K > 0$, Σ_t is a 3-sphere (resulting in a closed universe); when $K = 0$ or $K < 0$, Σ_t is the Euclidean space or a circular hyperboloid, respectively (resulting in an open universe).

10.1.3 The Robertson-Walker Metric

The cosmic spacetime should be equipped with a metric g_{ab} such that, on each surface Σ_t of homogeneity, the induced metric h_{ab} is one of those we obtained in Sect. 10.1.2 (corresponding to the line element dl^2). One can introduce a suitable coordinate system so that the line element of g_{ab} can be expressed in a simple form. To do that, we first point out the following conclusion: for two arbitrary isotropic observers A and B , and for two arbitrary surfaces Σ_{t_1} , Σ_{t_2} of homogeneity with $t_1 < t_2$, the world line segments of A and B between Σ_{t_1} and Σ_{t_2} are of the same length. Physically, it is easy to accept this statement, since all isotropic observers should be on an equal footing, and the existence of an exceptional observer is hard to imagine. In Optional Reading 10.1.4, we will give a rigorous proof for this conclusion.

Now let us introduce the coordinate system. On a surface Σ_0 of homogeneity, set (local) coordinates $x^1 \equiv \psi$, $x^2 \equiv \theta$ and $x^3 \equiv \varphi$ (as described in Sect. 10.1.2 for the different signs of K), then the world lines of isotropic observers can carry these coordinates out of Σ_0 in the following way: along the world line γ of an isotropic observer, the coordinates ψ , θ and φ remain constants, determined by their values at the intersecting point of γ and Σ_0 . Next, set the standard clock carried by each isotropic observer (i.e., the proper time τ) to zero on Σ_0 , and define the coordinate time t at each spacetime point p to be the proper time τ of the isotropic observer passing through p . In this way, we have a coordinate system $\{t, x^i\}$ of the cosmic spacetime, called the **Robertson-Walker (RW) coordinate system**. This system is obviously a comoving coordinate system of an isotropic reference frame. Note that the value of t on a surface of homogeneity can be different from the parameter for the family $\{\Sigma_t\}$ of the surfaces of homogeneity, since the parameter for the family in principle can be assigned arbitrarily. In other words, the homogeneous foliation $\{\Sigma_t\}$ can be associated with a function different than the coordinate t . However, to

to avoid confusion, we will stipulate that the homogeneous foliation $\{\Sigma_t\}$ is indeed associated with the coordinate t in the RW system. That is, for an arbitrary surface Σ_t of homogeneity, the time coordinate at every point equals the parameter of Σ_t in the foliation. The RW coordinate system has two major virtues:

① Each constant- t surface is a surface of homogeneity. Therefore a surface Σ_t of homogeneity is also a surface of simultaneity, representing the whole space of the universe at a time t .

② The world line of an isotropic observer is also a t -coordinate line, with its coordinate time being its proper time τ , called the **cosmic time**. Unless otherwise stated, the “time” in cosmology refers to the cosmic time.

Due to the virtue ②, the coordinate basis vector $(\partial/\partial t)^a$ equals the 4-velocity Z^a of isotropic observers. Hence

$$g_{00} = g_{ab}(\partial/\partial t)^a(\partial/\partial t)^b = g_{ab}Z^aZ^b = -1.$$

Due to the virtue ①, the three spatial coordinate basis vectors $(\partial/\partial x^i)^a$ are all tangent to surfaces of homogeneity, and thus are orthogonal to $(\partial/\partial t)^a$. Hence,

$$g_{0i} = g_{ab}(\partial/\partial t)^a(\partial/\partial x^i)^b = 0, \quad i = 1, 2, 3.$$

Since h_{ab} is the metric induced by g_{ab} , we have

$$g_{ij} = g_{ab}(\partial/\partial x^i)^a(\partial/\partial x^j)^b = h_{ab}(\partial/\partial x^i)^a(\partial/\partial x^j)^b = h_{ij}, \quad i, j = 1, 2, 3,$$

where the definition of the induced metric (see Definition 1 in Sect. 4.4) is used in the second step. Note that generally speaking, h_{ij} depends on t, x^1, x^2 and x^3 , we may denote it by $h_{ij}(t, x)$ (where x stands for x^1, x^2, x^3). By means of the uniqueness of the homogeneous foliation, it can be proved that (see Optional Reading 10.1.5) $h_{ij}(t, x)$ has the form of “separation of variables”, i.e.,

$$h_{ij}(t, x) = a^2(t)\hat{h}_{ij}(x), \tag{10.1.21}$$

with $a(t)$ depending only on t , and $\hat{h}_{ij}(x)$ depending only on x^i . Consequently, the line element of the cosmic metric g_{ab} in the RW coordinate system reads

$$ds^2 = -dt^2 + a^2(t)\hat{h}_{ij}(x)dx^i dx^j. \tag{10.1.22}$$

The induced line element on Σ_0 is

$$dl^2 = a^2(0)\hat{h}_{ij}(x)dx^i dx^j,$$

which belongs to the three cases summarized in Sect. 10.1.2. First we look at the simplest case, i.e., case (b), where the spatial metric is flat. In terms of the Cartesian coordinates x^i , the induced line element on Σ_0 is $a^2(0)\hat{h}_{ij}(x)dx^i dx^j = \delta_{ij}dx^i dx^j$. We may further take $a(0) = 1$ so that $\hat{h}_{ij}(x) = \delta_{ij}$. Then, (10.1.22) in this case

becomes

$$ds^2 = -dt^2 + a^2(t)(dx^2 + dy^2 + dz^2) \quad [\text{case (b)}]. \quad (10.1.23b)$$

In terms of the spherical coordinates ψ, θ, φ , where $\psi \equiv (x^2 + y^2 + z^2)^{1/2}$, the above line element reads

$$ds^2 = -dt^2 + a^2(t)[d\psi^2 + \psi^2(d\theta^2 + \sin^2 \theta d\varphi^2)] \quad [\text{case (b)}]. \quad (10.1.23b')$$

The form of the function $a(t)$ is determined by the Einstein field equation (see Sect. 10.2.3). Similarly, if the line element induced on Σ_0 has the form of (10.1.18) or (10.1.20), then we have, respectively

$$\begin{aligned} dl^2 &= a^2(0)\hat{h}_{ij}dx^i dx^j = \frac{1}{K}[d\psi^2 + \sin^2 \psi(d\theta^2 + \sin^2 \theta d\varphi^2)], \\ dl^2 &= a^2(0)\hat{h}_{ij}dx^i dx^j = -\frac{1}{K}[d\psi^2 + \sinh^2 \psi(d\theta^2 + \sin^2 \theta d\varphi^2)]. \end{aligned}$$

For these two cases, we may further take $a^2(0)K = 1$ and $a^2(0)K = -1$, respectively. Then, (10.1.22) in cases (a) and (c) becomes

$$ds^2 = -dt^2 + a^2(t)[d\psi^2 + \sin^2 \psi(d\theta^2 + \sin^2 \theta d\varphi^2)] \quad [\text{case (a)}], \quad (10.1.23a)$$

$$ds^2 = -dt^2 + a^2(t)[d\psi^2 + \sinh^2 \psi(d\theta^2 + \sin^2 \theta d\varphi^2)] \quad [\text{case (c)}]. \quad (10.1.23c)$$

Remark 1 The spacetime metric in (10.1.23b) is curved (unless $a(t)$ is a constant, while in Sect. 10.2.3 we will see that $a(t)$ is not a constant). It is each 3-dimensional surface of homogeneity that is flat, not the spacetime itself. Therefore, the spacetime corresponding to case (b) is a curved spacetime with a flat spatial geometry on each slice.

For cases (a), (b) and (c), we define r as follows:

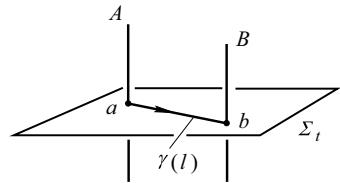
$$r = \begin{cases} \sin \psi, & [\text{case (a)}] \\ \psi, & [\text{case (b)}] \\ \sinh \psi, & [\text{case (c)}] \end{cases}, \quad (10.1.24)$$

Then, (10.1.23a), (10.1.23b') and (10.1.23c) can be combined to become

$$ds^2 = -dt^2 + a^2(t) \left[\frac{dr^2}{1 - kr^2} + r^2(d\theta^2 + \sin^2 \theta d\varphi^2) \right], \quad (10.1.25)$$

where

Fig. 10.6 Spacetime points a and b represent galaxies A and B at the time t . The length of the geodesic segment $\gamma(l)$ passing through them is the distance between galaxies A and B at t



$$k \equiv \begin{cases} 1, & [\text{case (a)}] \\ 0, & [\text{case (b)}] \\ -1, & [\text{case (c)}] \end{cases} \quad (\text{notice the similarity and distinction between } k \text{ and } K).$$

The metric in (10.1.23b) or (10.1.25) is called the **Robertson-Walker metric**,⁴ or **RW metric** for short. The above discussion indicates that, by only the cosmological principle (and the assumed uniqueness of the homogeneous foliation), the cosmic spacetime metric can be determined to be an RW metric, up to only two undetermined factors: ① the value of k specifying which case our universe belongs to (which will be discussed in Sect. 10.3.2); ② the function $a(t)$, determined by Einstein's equation (see Sect. 10.2.3).

Before the end of this section, let us talk about the physical meaning of $a(t)$. Suppose h_{ab} is the metric on a surface Σ_t of homogeneity induced by the RW metric. Then (Σ_t, h_{ab}) is a 3-dimensional Riemannian space. Let a and b be the intersecting points of Σ_t and two galaxies A and B , respectively, and $\gamma(l)$ be the geodesic segment lying on Σ_t connecting a and b (see Fig. 10.6). Then the arc length of $\gamma(l)$ is the distance between a and b in the space Σ_t . In physics, this length can be interpreted as the distance between the galaxies A and B at the time t , denoted by $D_{AB}(t)$. Let l_1 and l_2 be the parameters of $\gamma(l)$ at a and b (assuming $l_1 < l_2$), respectively. Then, according to the definition of arc length,

$$D_{AB}(t) = \int_{l_1}^{l_2} \sqrt{h_{ab} \left(\frac{\partial}{\partial l} \right)^a \left(\frac{\partial}{\partial l} \right)^b} dl.$$

Define $\hat{h}_{ab} \equiv a^{-2}(t) h_{ab}$ with $a(t) > 0$, then

$$D_{AB}(t) = a(t) \hat{D}_{AB}, \quad (10.1.26)$$

where

⁴ This metric is also called the Friedmann–Lemaître–Robertson–Walker (FLRW) metric, since A. Friedmann and G. Lemaître had considered certain special cases as dynamical solutions to Einstein's equation much earlier than H. P. Robertson and A. G. Walker (see Sect. 10.2.3). Whereas Robertson and Walker (each independently) obtained this general form of the metric first as purely a geometric result under the spatial homogeneity and isotropy conditions before using it to solve Einstein's equation.

$$\hat{D}_{AB} = \int_{l_1}^{l_2} \sqrt{\hat{h}_{ab} \left(\frac{\partial}{\partial l} \right)^a \left(\frac{\partial}{\partial l} \right)^b} dl. \quad (10.1.27)$$

From the above equation and (10.1.25), we can see that \hat{D}_{AB} is only determined by the galaxies A and B , and it is independent of time. Equation (10.1.26) then indicates that $a(t)$ is the value obtained by measuring the distance between A and B at t , with \hat{D}_{AB} as the unit. Hence, $a(t)$ reflects the time dependence of the distance between any two galaxies, and thus is called the **scale factor** of the universe. If the spatial coordinates of the galaxies A and B are (r_A, θ, φ) and (r_B, θ, φ) , then the parameter l can be chosen to be r along the geodesic. It can be verified that both θ and φ remain constants along the geodesic $\gamma(l)$. Then (10.1.26) can be expressed as

$$D_{AB}(t) = a(t) \int_{r_A}^{r_B} \frac{dr}{\sqrt{1 - kr^2}}. \quad (10.1.28)$$

It is easy to perform the above integral for all cases of $k = 1, 0$ and -1 . Note that the $-dt^2$ in (10.1.25) is $-c^2 dt^2$ in SI, which has the dimension of length. Thus, when $k = \pm 1$, the coordinate r is dimensionless, and $a(t)$ has the dimension of length; when $k = 0$, the dimensions of $a(t)$ and r can be arbitrary, with $a(t)r$ having the dimension of length.

When $k = 1$, the universe is closed. In this case, one can also ask about the volume of the universe at any time t (as the volume of a 3-sphere), which is obviously related to $a(t)$. It follows from (10.1.23a) that the volume element associated with the spatial induced metric h_{ab} is

$$\hat{\epsilon} = a^3 \sin^2 \psi \sin \theta d\psi \wedge d\theta \wedge d\varphi.$$

Hence, the volume of the whole space (as a 3-sphere), is

$$V = \int \hat{\epsilon} = a^3 \int_0^{2\pi} d\varphi \int_0^\pi \sin \theta d\theta \int_0^\pi \sin^2 \psi d\psi = 2\pi^2 a^3. \quad (10.1.29)$$

Therefore, $a^3(t)$ is proportional to the volume of the universe at the time t , and $a(t)$ is exactly the radius of the closed universe at t .

[Optional Reading 10.1.4]

Now we will prove a statement we claimed in the beginning of Sect. 10.1.3, which is summarized as Proposition 10.1.8. But before that, we shall introduce some facts which will be useful in the proof of Proposition 10.1.8 and the later discussions.

First, there is a theorem which assures that each point in a Riemannian space has a **convex neighborhood**, in which any pair of points can be joined by a unique geodesic segment lying in it [for a proof of this theorem, see Hicks (1965)]. Then, we have the following proposition.

Proposition 10.1.7 *Suppose (M, g_{ab}) is a spacetime satisfying the cosmological principle, which has a unique homogeneous foliation. Let Σ_t be a slice in the foliation, and P be*

an isotropic observer which intersects Σ_t at p , i.e., $p = P \cap \Sigma_t$.⁵ Suppose N is a convex neighborhood of p in Σ_t . Then, for an arbitrary isotropic observer P' such that $P' \cap \Sigma_t \subset N$, there exists an isometry $\psi : M \rightarrow M$ of g_{ab} satisfying the following conditions:

- (1) For each hypersurface $\Sigma_{t'}$ of homogeneity, $\psi[\Sigma_{t'}] = \Sigma_{t'}$.
- (2) Each isotropic observer is mapped by ψ to an isotropic observer;
- (3) P is mapped by ψ to P' ;
- (4) There exists an isotropic observer which is preserved under ψ (i.e., each point on the world line is fixed by ψ).

Proof Let $p' \in N$ be the intersecting point of P' and Σ_t . Since N is a convex neighborhood of p , there is a unique geodesic $\gamma(l)$ lying in N from p to p' , with l the arc length parameter. Let q be the middle point of $\gamma(l)$, i.e., $l_{pq} = l_{qp'}$, and let $w^a \equiv -w'^a = -(\partial/\partial l)^a|_q$. It follows from the definition of isotropy that there exists an isometry $\psi : M \rightarrow M$ such that $\psi(q) = q$, $\psi_*(Z^a|_q) = Z^a|_q$ and $\psi_*w^a = w'^a$, where Z^a is the 4-velocity of the isotropic observers. Then, (1) due to the uniqueness of the homogeneous foliation and Proposition 10.1.3, we have $\{\psi[\Sigma_{t'}]\} = \{\Sigma_{t'}\}$, and $\psi(q) = q$ assures that $\psi[\Sigma_{t'}] = \Sigma_{t'}$ for each hypersurface of homogeneity. (2) It follows from Corollary 10.1.4 that ψ sends one isotropic observer to another. (3) Since geodesics and arc lengths are all defined based on the metric g_{ab} , ψ being an isometry makes $\psi[\gamma_{pq}] = \gamma_{qp'}$. Thus, $\psi(p) = p'$, and so $\psi[P] = P'$. (4) Let Q be the isotropic observer passing through q . Since $\psi[Q]$ is still an isotropic observer, $\psi(q) = q$ indicates that $\psi[Q] = Q$. Then, for each point $q' \in Q$, $\psi(q') = q'$, i.e., Q is fixed pointwisely by ψ . \square

Proposition 10.1.8 Assume that the spacetime (M, g_{ab}) satisfies the cosmological principle and admits a unique homogeneous foliation. Then, for any pair of isotropic observers and any pair of surfaces Σ_{t_1} and Σ_{t_2} of homogeneity, the lengths of the world line segments of these two observers between Σ_{t_1} and Σ_{t_2} are equal.⁶

Proof Suppose A and B are two different isotropic observers, and Σ_{t_1} and Σ_{t_2} are two different surfaces of homogeneity. Let a_1 and a_2 be the intersection points of the world line of A and Σ_{t_1} and Σ_{t_2} , i.e., $a_1 = A \cap \Sigma_{t_1}$ and $a_2 = A \cap \Sigma_{t_2}$. Similarly, let $b_1 = B \cap \Sigma_{t_1}$ and $b_2 = B \cap \Sigma_{t_2}$ (Fig. 10.7).

Now being mindful that (Σ_{t_1}, h_{ab}) is a 3-dimensional Riemannian space, let us first suppose that b_1 is in a convex neighborhood of a_1 . Then, according to Proposition 10.1.7, there exists an isometry ψ of (M, g_{ab}) satisfying $\psi[\Sigma_{t_1}] = \Sigma_{t_1}$, $\psi[\Sigma_{t_2}] = \Sigma_{t_2}$ and $\psi[A] = B$. Since $a_2 = A \cap \Sigma_{t_2}$ and $b_2 = B \cap \Sigma_{t_2}$, it follows that $\psi(a_2) = b_2$. Hence, for the line segments $A_{a_1a_2}$ and $B_{b_1b_2}$, we have $\psi[A_{a_1a_2}] = B_{b_1b_2}$. As a consequence, the arc lengths of $A_{a_1a_2}$ and $B_{b_1b_2}$ are equal.

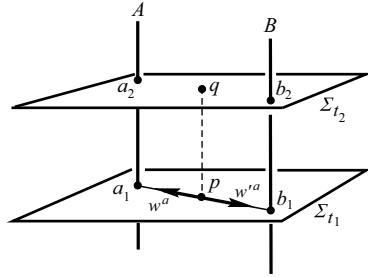
Since (Σ_{t_1}, h_{ab}) is connected, and since it can be covered by convex neighborhoods, One can easily show that the arc lengths of $A_{a_1a_2}$ and $B_{b_1b_2}$ are still equal even if b_1 is not in a convex neighborhood of a_1 . \square

[The End of Optional Reading 10.1.4]

⁵ Technically, $P \cap \Sigma_t = \{p\}$ is a set with p as its only element. We will recognize it as a point for convenience.

⁶ If the homogeneous foliation is not unique, then the lengths of the world line segments of two isotropic observers between Σ_{t_1} and Σ_{t_2} can be different, even if Σ_{t_1} and Σ_{t_2} are leaves in the same homogeneous foliation.

Fig. 10.7 It takes isotropic observers A and B the same proper time from Σ_{t_1} to Σ_{t_2}



[Optional Reading 10.1.5]

Now we provide the proof for (10.1.21). That is, $h_{ij}(t, x)$ has the form of “separation of variables”.

Suppose Σ_{t_1} and Σ_{t_2} are surfaces of homogeneity, and G is an isotropic observer. Let $p_1 \equiv G \cap \Sigma_{t_1}$ and $p_2 \equiv G \cap \Sigma_{t_2}$, which can be expressed in terms of coordinates as $p_1 = (t_1, x_G^i)$ and $p_2 = (t_2, x_G^i)$. Let $X^a|_{p_1}$ and $Y^a|_{p_1}$ be two spatial vectors at p_1 with the same magnitude, whose coordinate components are X^i and Y^i , respectively, i.e., $X^a|_{p_1} = X^i(\partial/\partial x^i)^a|_{p_1}$ and $Y^a|_{p_1} = Y^i(\partial/\partial x^i)^a|_{p_1}$. Then

$$h_{ij}(t_1, x_G) X^i X^j = h_{ij}(t_1, x_G) Y^i Y^j. \quad (10.1.30)$$

Since G is an isotropic observer, there is an isometry $\psi : M \rightarrow M$ such that $\psi(p_1) = p_1$ and $\psi_*(X^a|_{p_1}) = Y^a|_{p_1}$. From the definition of the RW coordinate system and Corollary 10.1.4, we can see that the coordinate transformation $\{t, x^i\} \mapsto \{t', x'^i\}$ induced by ψ satisfies

$$t' = t, \quad x'^i = f^i(x), \quad i = 1, 2, 3, \quad (10.1.31)$$

and $x_G^i = f^i(x_G)$, where the x in the parentheses stands for x^1, x^2 and x^3 , and similarly for x_G . It follows from (4.1.7) that

$$Y^i = X^j \frac{\partial f^i}{\partial x^j} \Big|_{x_G}. \quad (10.1.32)$$

Consider the spatial vectors $X^a|_{p_2} = X^i(\partial/\partial x^i)^a|_{p_2}$ and $Y^a|_{p_2} = Y^i(\partial/\partial x^i)^a|_{p_2}$ at p_2 . Since the real numbers X^i and Y^i satisfy (10.1.32), we have $\psi_*(X^a|_{p_2}) = Y^a|_{p_2}$. Hence, $X^a|_{p_2}$ and $Y^a|_{p_2}$ are of the same magnitude. That is,

$$h_{ij}(t_2, x_G) X^i X^j = h_{ij}(t_2, x_G) Y^i Y^j. \quad (10.1.33)$$

Next, suppose $X^a|_{p_1}$ and $Y^a|_{p_1}$ do not have the same magnitude (and $Y^a|_{p_1} \neq 0$). Since the 3×3 matrix constituted by h_{ij} is positive definite, there exists $\lambda \in \mathbb{R}$ such that

$$h_{ij}(t_1, x_G) X^i X^j = \lambda^2 h_{ij}(t_1, x_G) Y^i Y^j = h_{ij}(t_1, x_G)(\lambda Y^i)(\lambda Y^j),$$

and hence

$$h_{ij}(t_2, x_G) X^i X^j = h_{ij}(t_2, x_G)(\lambda Y^i)(\lambda Y^j) = \lambda^2 h_{ij}(t_2, x_G) Y^i Y^j.$$

Therefore, for any nonvanishing row vectors (X^1, X^2, X^3) and (Y^1, Y^2, Y^3) , we have

$$\frac{h_{ij}(t_1, x_G) X^i X^j}{h_{kl}(t_2, x_G) X^k X^l} = \frac{h_{ij}(t_1, x_G) Y^i Y^j}{h_{kl}(t_2, x_G) Y^k Y^l}. \quad (10.1.34)$$

Note that the indices i, j, k and l in the above equation are all summed over 1, 2 and 3. The ratio in the above equation does not depend on (X^1, X^2, X^3) and (Y^1, Y^2, Y^3) , but is only determined by t_1, t_2 and x_G . Thus, there exists a real number $\omega(t_1, t_2, x_G)$ such that

$$h_{ij}(t_1, x_G)X^i X^j = \omega(t_1, t_2, x_G) h_{ij}(t_2, x_G)X^i X^j, \quad \forall X^1, X^2, X^3 \in \mathbb{R}. \quad (10.1.35)$$

Consequently, for arbitrary (X^1, X^2, X^3) and (Y^1, Y^2, Y^3) , we have

$$\begin{aligned} h_{ij}(t_1, x_G)(X^i + Y^i)(X^j + Y^j) &= \omega(t_1, t_2, x_G) h_{ij}(t_2, x_G)(X^i + Y^i)(X^j + Y^j), \\ h_{ij}(t_1, x_G)(X^i - Y^i)(X^j - Y^j) &= \omega(t_1, t_2, x_G) h_{ij}(t_2, x_G)(X^i - Y^i)(X^j - Y^j), \end{aligned}$$

and thus for arbitrary (X^1, X^2, X^3) and (Y^1, Y^2, Y^3) ,

$$h_{ij}(t_1, x_G)X^i Y^j = \omega(t_1, t_2, x_G) h_{ij}(t_2, x_G)X^i Y^j.$$

Hence, $h_{ij}(t_1, x_G) = \omega(t_1, t_2, x_G)h_{ij}(t_2, x_G)$. Since the isotropic observer G is arbitrary, we can simply denote x_G as x and obtain

$$h_{ij}(t_1, x) = \omega(t_1, t_2, x)h_{ij}(t_2, x). \quad (10.1.36)$$

Finally, now we show that ω actually does not depend on x . For an arbitrary isotropic observer G with $p_1 = G \cap \Sigma_{t_1}$ and $p_2 = G \cap \Sigma_{t_2}$, there is a convex neighborhood N of p_1 in Σ_{t_1} . For simplicity, we may assume that the coordinate patch of x^i covers N . Then, according to Proposition 10.1.7, for an arbitrary point $p'_1 \in N$, there is an isometry ϕ of (M, g_{ab}) satisfying ① ϕ maps each Σ_t to itself; ② ϕ maps each isotropic observer to an isotropic observer, and, especially, ③ ϕ maps G to G' , the isotropic observer that contains $p'_1 = \phi(p_1) \in G'$. Let $p'_2 \in G' \cap \Sigma_{t_2}$, then it follows that $p'_2 = \phi(p_2)$. Because of ① and ② above, the coordinate transformation induced by ϕ will have the form of (10.1.31). In terms of the new coordinates t' and x'^i , the line element of the cosmic metric g_{ab} can be expressed as

$$ds^2 = -dt'^2 + h_{ij}(t', x')dx'^i dx'^j = -dt^2 + h_{ij}(t, x')\frac{\partial f^i}{\partial x^k} \frac{\partial f^j}{\partial x^l} dx^k dx^l,$$

Since ϕ is an isometry, comparing with the line element in the old coordinate system $ds^2 = -dt^2 + h_{kl}(t, x)dx^k dx^l$, we obtain

$$h_{ij}(t, x')\frac{\partial f^i}{\partial x^k} \frac{\partial f^j}{\partial x^l} = h_{kl}(t, x). \quad (10.1.37)$$

Setting t to be t_1 and t_2 yields

$$h_{ij}(t_1, x')\frac{\partial f^i}{\partial x^k} \frac{\partial f^j}{\partial x^l} = h_{kl}(t_1, x), \quad (10.1.38)$$

$$h_{ij}(t_2, x')\frac{\partial f^i}{\partial x^k} \frac{\partial f^j}{\partial x^l} = h_{kl}(t_2, x). \quad (10.1.39)$$

Applying (10.1.36) to both sides of (10.1.38), we have

$$\omega(t_1, t_2, x')h_{ij}(t_2, x')\frac{\partial f^i}{\partial x^k} \frac{\partial f^j}{\partial x^l} = \omega(t_1, t_2, x)h_{kl}(t_2, x).$$

Noticing (10.1.39), we obtain $\omega(t_1, t_2, x') = \omega(t_1, t_2, x)$. To see more clearly that ω does not depend on x , let us go to the active perspective, i.e., viewing the coordinate transformation $x^i \rightarrow x'^i$ under ψ as the map between two observers G and G' in the old coordinates x^i . Then, we have

$$\omega(t_1, t_2, x_{G'}) = \omega(t_1, t_2, x_G).$$

Since G' is arbitrary, the above equation shows that $\omega(t_1, t_2, x_G)$ is independent of x_G for $G \cap N \neq \emptyset$. As Σ_{t_1} can be covered by convex neighborhoods, it follows that $\omega(t_1, t_2, x)$ does not depend on x , so it can be denoted by $\omega(t_1, t_2)$. Then, (10.1.36) turns out to be $h_{ij}(t_1, x) = \omega(t_1, t_2)h_{ij}(t_2, x)$. Particularly, let t_2 be fixed and t_1 be arbitrary. Denoting t_1 by t , we have

$$h_{ij}(t, x) = \omega(t, t_2)h_{ij}(t_2, x). \quad (10.1.40)$$

Let $\hat{h}_{ij}(x) \equiv h_{ij}(t_2, x)$ and $a^2(t) \equiv \omega(t, t_2)$, the above equation becomes $h_{ij}(t, x) = a^2(t)\hat{h}_{ij}(x)$, i.e., (10.1.21).

[The End of Optional Reading 10.1.5]

10.2 Dynamics of the Universe

10.2.1 The Hubble-Lemaître Law

In the early 20th century, the American astronomer V. S. Slipher observed the spectral lines of 41 galaxies. He discovered redshifts within 36 of these galaxies. Recall that a **redshift** is defined to be $z \equiv (\lambda' - \lambda)/\lambda$, where λ and λ' are the wave lengths of light when it is emitted and observed, respectively. Attributing the redshifts to the Doppler effect, Slipher's discovery shows that these 36 galaxies are moving away from our galaxy, the Milky Way. In other words, this indicates that our universe is expanding. (Since the solar system is orbiting around the Galactic Center, i.e., the center of the Milky Way, the blueshifts of the other 5 galaxies can be interpreted as being caused by their motion toward the Sun.) So far, spectra from tens of thousand of galaxies have been measured, and all of them are redshift except for few (those from nearby galaxies). This provides a solid observational basis for the expansion of the universe. In 1923, the American astronomer Edwin Hubble began to make measurements of the distance of the extragalactic galaxies from us, which is more difficult than the measurement of redshifts. He found that the redshift z of a galaxy is proportional to its distance D from us, and z is equal to the recessional speed u when the latter is very small. [The Taylor expansion of (6.6.66a) to the first order yields $z \cong u$.] Hence, Hubble published the well-known **Hubble law** in 1929 stating that,

$$u_0 = H_0 D_0, \quad (\text{the subscript 0 stands for "the current value"}) \quad (10.2.1)$$

where H_0 is a constant independent of galaxies, known as the **Hubble constant**.

From the theory side, it is not difficult to derive Hubble's law from the RW metric. Define the relative speed, also called the speed of separation, between two galaxies as $u(t) := dD(t)/dt$, where $D(t)$ is the proper distance between them at a time t . From (10.1.26) we can easily see that,

$$u(t) = \hat{D} \frac{da(t)}{dt} = \frac{\dot{a}(t)}{a(t)} D(t), \quad \dot{a}(t) \equiv \frac{da(t)}{dt}. \quad (10.2.2)$$

Thus, one can define the **Hubble parameter**

$$H(t) := \frac{\dot{a}(t)}{a(t)}, \quad (10.2.3)$$

which is independent of both the galaxies and the distance between the galaxies, so that

$$u(t) = H(t)D(t). \quad (10.2.4)$$

The above equation indicates that, at any time t , the recessional speed (the speed of separation) between two galaxies is proportional to the proper distance between them. Let t_0 be the present time, and denote $H(t_0)$ simply by H_0 (i.e., the Hubble constant), then

$$u(t_0) = H_0 D(t_0), \quad (10.2.1')$$

which is exactly (10.2.1). The result in (10.2.4) was derived by the Belgian physicist G. Lemaître two years before Hubble's article, and thus more properly, Hubble's law is also called the **Hubble-Lemaître law**. Note that the Hubble parameter is different from the Hubble constant in that the former depends on t , while the latter is merely the current value of the former. Because observational results indicate that $H_0 > 0$, it follows from (10.2.1') that $u(t_0) > 0$ whenever $D(t_0) \neq 0$. That is, on an arbitrary galaxy, the observation for another galaxy will show that the latter is moving away. [The measurement by Hubble only reveals that other galaxies are going away from the Milky Way, while (10.2.1) asserts that any pair of galaxies are going away from each other.] Thus, the fact that all galaxies are measured to be away from us does not mean that the Milky Way is the center of the expanding universe. As an analogy, one can imagine a balloon with lots of ants on its surface. When such a balloon is expanding, each of these ants finds that the others are going away from it, with no ants being more special than the others. Just like the expanding balloon, there is no center of expansion in the universe.

According to (10.2.1), u could be greater than the speed of light in vacuum when D is large enough. This does not contradict relativity. To see this, recall that one of the principles of relativity states that “the world line of a point mass must be timelike.” This is an absolute and unambiguous statement, which, by a properly defined concept of speed, is equivalent to the statement that “the speed of a point mass is less than the speed of light in vacuum” (which is a relative statement). One must keep in mind that the “speed” in the latter statement refers to the magnitude of the 3-velocity u^a defined in (6.3.28), i.e., the 3-speed of a point mass obtained from a local measurement by an instantaneous observer. If the observer is an inertial observer in Minkowski spacetime, then the speed is nothing but the speed of the

particle relative to the inertial frame the observer belongs to. However, there are various definitions of speed. For a definition different from the above definition of speed, a speed greater than the speed of light does not necessarily violate relativity. The recessional speed of galaxies is such an example. It is defined as the derivative of the distance of the galaxies with respect to the cosmic time, which is, of course, of physical meaningful and reasonable to be called a speed. However, this is not the speed obtained from a local measurement by an instantaneous observer, and hence it is not a contradiction to the principle of relativity. In fact, when deriving the RW metric, the world line of each galaxy has been recognized to be timelike, which automatically obeys the principle of relativity stated above. As a consequence, for any instantaneous observer (not necessarily an isotropic observer), the speed of a galaxy obtained by a local measurement is certainly less than the speed of light in vacuum.

10.2.2 Cosmological Redshift

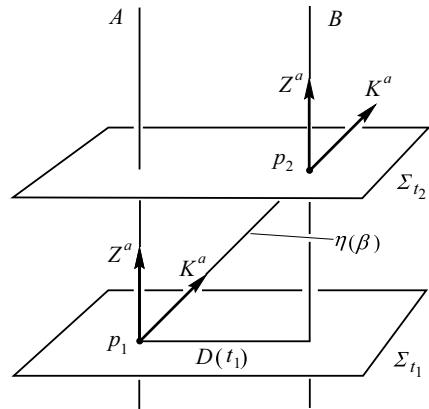
Hubble interpreted cosmic redshifts as the Doppler effect in flat spacetime, from which he obtained the recessional speed of the galaxies. According to general relativity, the existence of matter in the universe results in the curvature of spacetime, and cosmic redshifts are actually an effect of the curved spacetime. Compared to cosmological scales, the galaxies Hubble measured are of very small distances from us, and the redshift is relatively small. In this case, it is acceptable to treat these redshifts as the Doppler effect. However, for galaxies at sufficiently large distances from us, their redshifts have to be interpreted as being due to the curved spacetime geometry.

Under the geometric optics approximation (see Optional Reading 7.2.1), light signals are regarded as propagating along null geodesics. Suppose a photon emitted by a galaxy A at p_1 travels along a null geodesic $\eta(\beta)$ (where β is an affine parameter), and this photon is received by another galaxy B at p_2 (see Fig. 10.8). Let $K^a = (\partial/\partial\beta)^a$ be the wave 4-vector of the above photon, then its angular frequency at p_1 relative to the observer A is $\omega_1 = -g_{ab}Z^aK^b|_{p_1}$, where $Z^a|_{p_1}$ is the 4-velocity of A at p_1 . Noticing that Z^a is the same as the coordinate basis vector $(\partial/\partial t)^a$ in the RW coordinate system $\{t, r, \theta, \varphi\}$, and that

$$K^b = \frac{dt}{d\beta} \left(\frac{\partial}{\partial t} \right)^b + \frac{dx^i}{d\beta} \left(\frac{\partial}{\partial x^i} \right)^b,$$

we have $\omega_1 = dt/d\beta|_{p_1}$. Similarly, the angular frequency of this photon at p_2 relative to the observer B is $\omega_2 = dt/d\beta|_{p_2}$. The cosmic redshift can be obtained by means of the geodesic equation of $\eta(\beta)$:

Fig. 10.8 Derivation of the cosmic redshift. $\eta(\beta)$ is a null geodesic



$$\frac{d^2x^\mu}{d\beta^2} + \Gamma^\mu_{\nu\sigma} \frac{dx^\nu}{d\beta} \frac{dx^\sigma}{d\beta} = 0, \quad (\mu = 0, 1, 2, 3), \quad (10.2.5)$$

where x^0, x^1, x^2, x^3 are respectively t, r, θ, φ . From (10.1.25), we can obtain the Christoffel symbols, among which the nonvanishing ones are

$$\begin{aligned} \Gamma^0_{11} &= \frac{a\dot{a}}{1 - kr^2}, & \Gamma^0_{22} &= a\dot{a}r^2, & \Gamma^0_{33} &= a\dot{a}r^2 \sin^2 \theta, \\ \Gamma^1_{01} &= \Gamma^1_{10} = \frac{\dot{a}}{a}, & \Gamma^1_{11} &= \frac{kr}{1 - kr^2}, \\ \Gamma^1_{22} &= -r(1 - kr^2), & \Gamma^1_{33} &= -r(1 - kr^2) \sin^2 \theta, \\ \Gamma^2_{02} &= \Gamma^2_{20} = \Gamma^3_{03} = \Gamma^3_{30} = \frac{\dot{a}}{a}, & \Gamma^2_{12} &= \Gamma^2_{21} = \Gamma^3_{13} = \Gamma^3_{31} = \frac{1}{r}, \\ \Gamma^2_{33} &= -\sin \theta \cos \theta, & \Gamma^3_{23} &= \Gamma^3_{32} = \cot \theta. \end{aligned}$$

It is easy to verify that the world line of an isotropic observer is a geodesic. We leave this as Exercise 10.2. [From the above expressions for the Christoffel symbols as well as (5.7.2), this is in fact almost obvious.] Setting $\mu = 2, 3$ in (10.2.5), we have

$$\begin{aligned} \frac{d^2\theta}{d\beta^2} + \frac{2\dot{a}}{a} \frac{dt}{d\beta} \frac{d\theta}{d\beta} + \frac{2}{r} \frac{dr}{d\beta} \frac{d\theta}{d\beta} - \left(\frac{d\varphi}{d\beta} \right)^2 \sin \theta \cos \theta &= 0, \\ \frac{d^2\varphi}{d\beta^2} + \frac{2\dot{a}}{a} \frac{dt}{d\beta} \frac{d\varphi}{d\beta} + \frac{2}{r} \frac{dr}{d\beta} \frac{d\varphi}{d\beta} + 2 \frac{d\theta}{d\beta} \frac{d\varphi}{d\beta} \cot \theta &= 0. \end{aligned} \quad (10.2.6)$$

The RW coordinate system can be chosen such that $\theta(p_1) = \theta_0$, $\varphi(p_1) = \varphi_0$ and that both the θ - and φ -components of $K^a|_{p_1}$ vanish. In this way, we have $\theta = \theta_0$ and $\varphi = \varphi_0$ along the whole geodesic $\eta(\beta)$. That is, for an arbitrary geodesic $\eta(\beta)$, one can always redefine θ and φ and make $\eta(\beta)$ a radial geodesic. [Since $\eta(\beta)$ has

been given, functions $t(\beta)$, $r(\beta)$, $\theta(\beta)$ and $\varphi(\beta)$ in its parametric equation are all determined in a given coordinate system. To show that $\theta(\beta) = \theta_0$ and $\varphi(\beta) = \varphi_0$, one only needs to notice that (10.2.6) is a system of 2nd-order equations for two unknown functions $\theta(\beta)$ and $\varphi(\beta)$, while $\theta(\beta) = \theta_0$ and $\varphi(\beta) = \varphi_0$ give the unique solution satisfying the initial conditions $\theta(p_1) = \theta_0$, $\varphi(p_1) = \varphi_0$, $\frac{d\theta}{d\beta}|_{p_1} = 0$ and $\frac{d\varphi}{d\beta}|_{p_1} = 0$.] Furthermore, setting $\mu = 0$ in (10.2.5), we have

$$\frac{d^2t}{d\beta^2} + \frac{a\dot{a}}{1-kr^2} \left(\frac{dr}{d\beta} \right)^2 = 0.$$

Since K^a is null, i.e., $K^a K^b g_{ab} = 0$, we also have

$$\left(\frac{dt}{d\beta} \right)^2 = \frac{a^2}{1-kr^2} \left(\frac{dr}{d\beta} \right)^2. \quad (10.2.7)$$

Combining the above two equations yields

$$\frac{d^2t}{d\beta^2} + \frac{\dot{a}}{a} \left(\frac{dt}{d\beta} \right)^2 = 0.$$

Define $\omega = dt/d\beta$, then the above equation becomes $\frac{d\omega}{d\beta} + \frac{\omega}{a} \frac{da}{d\beta} = 0$. Its general solution gives

$$\omega = \frac{\omega_0}{a}, \quad (10.2.8)$$

where ω_0 is a constant of integration. In the manner as we have discussed above, we see that for any value of β , the corresponding value of ω is the angular frequency of the photon measured by the isotropic observer that passes through the point $\eta(\beta)$. Thus, the above equation can be interpreted as follows: as the universe is expanding, the wavelength of each photon in the universe (with respect an isotropic observer) is stretched proportionally, which leads to the redshift. Applying (10.2.8) to points p_1 and p_2 , respectively, we obtain

$$\frac{\omega_2}{\omega_1} = \frac{a(t_1)}{a(t_2)}, \quad (10.2.9)$$

where $t_1 = t(p_1)$ and $t_2 = t(p_2)$. Hence, the relative redshift is

$$z = \frac{\lambda_2 - \lambda_1}{\lambda_1} = \frac{\omega_1}{\omega_2} - 1 = \frac{a(t_2)}{a(t_1)} - 1. \quad (10.2.10)$$

If the distance between A and B is sufficiently small, we have $t_2 - t_1 \cong D(t_1)$ because the world line $\eta(\beta)$ of the photon is null, as shown in Fig. 10.8. Neglecting the higher order terms in the Taylor expansion yields

$$a(t_2) \cong a(t_1) + \dot{a}(t_1)(t_2 - t_1) \cong a(t_1) + \dot{a}(t_1) D(t_1),$$

Hence,

$$z = \frac{\dot{a}(t_1)}{a(t_1)} D(t_1) = H(t_1) D(t_1), \quad (10.2.11)$$

where (10.2.3) is used in the last step. Denote $t_2 (\cong t_1)$ as t_0 , then the above equation is exactly the observational result by Hubble.

Remark 1 In Exercise 10.3 we will see another approach for deriving (10.2.8), which takes the advantage of the geodesic equation in the form of $K^a \nabla_a K^b = 0$, instead of using the component form (10.2.5). Alternatively, (10.2.8) can also be derived in a purely geometric fashion (which makes use of the fact that the contraction of the tangent vector field of a geodesic and a Killing field remains constant along the geodesic). For details, see Wald (1984), pp. 103–104.

10.2.3 Evolution of the Scale Factor

The Einstein tensor G_{ab} for the Robertson-Walker metric can be expressed in terms of $a(t)$. When T_{ab} , the energy-momentum tensor of all the content in the universe, is also expressed in terms of $a(t)$, Einstein's equation $G_{ab} = 8\pi T_{ab}$ will give rise to a set of differential equations for $a(t)$, from which we can solve for the time evolution of the universe.

The contents of the universe can be classified into two types: those consisting of particles with nonzero rest masses are called **matter**; those consisting of particles with zero rest mass are called **radiation**. Matter is mainly accumulated in galaxies, while the main contribution to radiation is the **cosmic microwave background radiation** (CMB, or CMBR), which is some electromagnetic microwaves distributed throughout the whole universe, discovered in 1965 (for details, see Sect. 10.3.1). On a cosmological scale, each galaxy can be treated as a point mass (like a drop in the ocean), and all the galaxies are regarded as forming a perfect fluid. The pressure of such a perfect fluid is negligible (namely the random motions of the galaxies can be neglected), and thus such a perfect fluid can be approximated as a dust, with each galaxy being a particle in this dust. Hence, the world line of each galaxy is a geodesic (see Sect. 6.5). Furthermore, since a perfect fluid is isotropic, each galaxy can be approximately regarded as an isotropic observer [as we have seen below (10.2.5), the world lines of isotropic observers are indeed geodesics]. The energy-momentum tensor of all the matter (i.e., the dust) in the universe can be expressed as

$$T_{ab}(\text{matter}) = \rho_M U_a U_b,$$

where U^a is the 4-velocity field of the isotropic observers, and ρ_M is the energy density of matter measured by the isotropic observers. On the other hand, all the radiation in the universe may also be treated as a special kind of perfect fluid, whose 4-velocity is the same as U^a . Then, the energy-momentum tensor of all the radiation in the universe reads

$$T_{ab}(\text{radiation}) = \rho_R U_a U_b + p (g_{ab} + U_a U_b),$$

where the energy density ρ_R and the pressure p of the radiation are both measured by the isotropic observers and satisfy $p = \rho_R/3$ [see (6.5.3)]. Combining these two contributions, the total energy-momentum tensor of the universe can be approximately written as

$$T_{ab} = \rho U_a U_b + p (g_{ab} + U_a U_b), \quad (10.2.12)$$

where $\rho = \rho_M + \rho_R$ is the sum of the energy densities of the dust (galaxies) and the radiation. In the actual universe, there are also other kinds of matter, in addition to galaxies. However, according to the cosmological principle, one can expect that their 4-velocities on average is still U^a , and so their energy-momentum tensors will also have the form of (10.2.12). In other words, the T_{ab} in (10.2.12) can be regarded as including the contributions from all kinds of matter in the universe (ρ and p have included the contributions from all of them). In summary, in the standard model, there are only two types of content in the universe, matter with the characteristic $p \cong 0$, and radiation with the characteristic $p = \rho/3$. The contributions from both of them have been included in (10.2.12), where ρ and p are independent of the spatial coordinates due to the spatial homogeneity of the universe.

In the RW line element (10.1.25), t, r, θ and φ are the comoving coordinates. The nonvanishing components of T_{ab} obtained from (10.2.12) in this system are

$$T_{00} = \rho, \quad T_{ij} = p g_{ij}, \quad (10.2.13)$$

where the only nonvanishing g_{ij} are

$$g_{11} = \frac{a^2}{1 - kr^2}, \quad g_{22} = a^2 r^2, \quad g_{33} = a^2 r^2 \sin^2 \theta.$$

On the other hand, from (10.1.25), one can find the nonvanishing components of the Einstein tensor G_{ab} (the calculation is left as an exercise), which can be expressed in terms of $a(t)$ as

$$G_{00} = \frac{3(\dot{a}^2 + k)}{a^2}, \quad (10.2.14)$$

$$G_{ij} = -\left(\frac{2\ddot{a}}{a} + \frac{\dot{a}^2 + k}{a^2}\right)g_{ij}. \quad (10.2.15)$$

Then, the components of Einstein's equation, $G_{00} = 8\pi T_{00}$ and $G_{ij} = 8\pi T_{ij}$, can be expressed as

$$\frac{3(\dot{a}^2 + k)}{a^2} = 8\pi\rho, \quad (10.2.16)$$

$$2\ddot{a} + \frac{\dot{a}^2 + k}{a^2} = -8\pi p. \quad (10.2.17)$$

Equations (10.2.16) and (10.2.17) are the fundamental equations that determine the scale factor $a(t)$. From these two equations we can easily get

$$\frac{\ddot{a}}{a} = -\frac{4\pi}{3}(\rho + 3p). \quad (10.2.18)$$

Equation (10.2.16) is called the **Friedmann equation**. Equations (10.2.16) and (10.2.18) are also called the **first Friedmann equation** and the **second Friedmann equation**, respectively. Differentiating (10.2.16) yields

$$\frac{\dot{a}}{a} \left(\frac{\ddot{a}}{a} - \frac{\dot{a}^2 + k}{a^2} \right) = \frac{4\pi\dot{\rho}}{3}. \quad (10.2.19)$$

Then, using the Friedmann equations (10.2.16) and (10.2.18) to eliminate \ddot{a}/a and $(\dot{a}^2 + k)/a^2$ in the above equation, we obtain

$$\dot{\rho} + 3(\rho + p)\frac{\dot{a}}{a} = 0. \quad (10.2.20)$$

On the other hand, once we have (10.2.16) and (10.2.20), we can apply both of them to (10.2.19) and get

$$\frac{\dot{a}}{a} \left(\frac{\ddot{a}}{a} + \frac{4\pi}{3}(\rho + 3p) \right) = 0.$$

Therefore, the Friedmann equations are equivalent to (10.2.16) and (10.2.20) when $\dot{a} \neq 0$.

For $\rho > 0$ and $p \geq 0$, (10.2.18) indicates that $\ddot{a} < 0$. This means that the universe is either expanding ($\dot{a} > 0$) or contracting ($\dot{a} < 0$), but cannot be static, since $\dot{a} = 0$ can at most happen at a special moment when $\dot{a} > 0$ is turning into $\dot{a} < 0$. Since the observation results indicate that the universe is expanding in the present day, i.e., $\dot{a}(t_0) > 0$ (t_0 is the current time coordinate), it follows from $\ddot{a} < 0$ that $\dot{a}(t) > \dot{a}(t_0) > 0$ for arbitrary $t < t_0$, and, the smaller t is, the greater $\dot{a}(t)$ is. Hence, when we trace backward in time, the universe shrinks more and more rapidly, and finally, at a certain time (set as $t = 0$), the value of a becomes zero. At this time, the density becomes infinite, and so we say that the universe expanded out of a singularity, called the **big bang singularity**. In fact, it is not so appropriate to refer to the origin of the

universe as a “big bang”. The word “bang” usually means a hit striking violently with a loud noise, which occurs as an event in a regular spacetime background, with no singularity at the spacetime point where the hit occurs. Furthermore, there exists something (e.g., a bomb before its explosion) whose world line ends in the future direction at the spacetime point where the bang occurs (each bomb fragment has its own world line after the explosion). The big bang of the universe is quite different. First, it corresponds to a spacetime singularity. All timelike geodesics are incomplete in the past direction, and all of them approach the singularity as t tends to zero: for any pair of such geodesics, $\gamma_1(t)$ and $\gamma_2(t)$, the distance between the spacetime points $\gamma_1(t)$ and $\gamma_2(t)$ tends to zero as $t > 0$ tends to zero. On the other hand, there does not exist a timelike geodesic that approaches the big bang singularity in the future direction. Intuitively, one may imagine that at the beginning of time, all particles in the universe are jammed in a spatial volume that “cannot be smaller”. During the expansion of the universe, each particle runs away from the others.

Before solving (10.2.16) and (10.2.20) in the general cases, let us discuss two extreme cases: ① the **dust-only universe**, whose contribution to T_{ab} comes completely from matter (dust); and ② the **radiation-only universe**, whose contribution to T_{ab} comes completely from radiation. For the dust-only universe, $p = 0$, and integration of (10.2.20) gives

$$\rho_M a^3 = \text{constant}. \quad (10.2.21)$$

This is pretty natural, because a comoving volume is proportional to a^3 , and the number of particles within this volume is a constant. Thus, the energy contained in this comoving volume is a constant, and so its density is proportional to a^{-3} . For the radiation-only universe, $p = \rho_R/3$, and integration of (10.2.20) gives

$$\rho_R a^4 = \text{constant}. \quad (10.2.22)$$

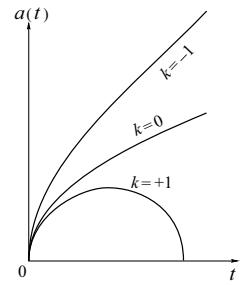
This is because the number of photons within a comoving volume is a constant, while the frequency (energy) of every photon is proportional to a^{-1} due to the redshift [see (10.2.8)]. Thus, the energy density of radiation is proportional to a^{-4} . Our present universe is matter-dominated, which is closer to a dust-only universe than to a radiation-only universe. However, when t is sufficiently small, the universe is radiation-dominated (although there were no galaxies yet in the early universe, only particles). In the following, we will solve (10.2.16) and (10.2.20) for these two extreme cases.

For the radiation-only universe, we write (10.2.22) as

$$B^2 = \frac{8\pi}{3} \rho a^4, \quad (10.2.23)$$

where $B > 0$ is a constant, and ρ_R is denoted by ρ . Then (10.2.16) can be rewritten as

Fig. 10.9 The curves of $a(t)$ for the radiation-only universe



$$\dot{a}^2 = \frac{B^2}{a^2} - k. \quad (10.2.24)$$

By setting $b(t) \equiv a^2(t)$, under the condition $a(t) = 0$ when $t \rightarrow 0$, we can find that a special solution of the above equation is $a^2(t) = 2Bt - kt^2$. Thus, for the three cases of k , we have

$$\text{case (a)} \quad (k = +1), \quad a^2(t) = 2Bt - t^2, \quad (10.2.25a)$$

$$\text{case (b)} \quad (k = 0), \quad a^2(t) = 2Bt, \quad (10.2.25b)$$

$$\text{case (c)} \quad (k = -1), \quad a^2(t) = 2Bt + t^2. \quad (10.2.25c)$$

The diagrams for $a(t)$ in these cases are shown in Fig. 10.9. Since radiation is dominant when t is sufficiently small, the behaviors of the three curves of $a(t)$ are of significance near the origin. It follows from (10.2.25) or Fig. 10.9 that $a = 0$ when $t = 0$, which corresponds to the big bang singularity. In (10.2.24), k can be neglected when a is sufficiently small. Therefore, the three curves are approximately the same near the origin.

For the matter (dust)-only universe, we write (10.2.21) as

$$A = \frac{8\pi}{3}\rho a^3, \quad (10.2.26)$$

where $A > 0$ is a constant, and ρ_M in (10.2.21) is replaced by ρ . Then, (10.2.16) can be rewritten as

$$\dot{a}^2 = \frac{A}{a} - k. \quad (10.2.27)$$

In order to solve it, we introduce a new variable

$$\hat{t}(t) \equiv \int_0^t \frac{dt'}{a(t')} . \quad (10.2.28)$$

Denote $da/d\hat{t}$ as a' , then (10.2.27) becomes

$$a'^2 = Aa - ka^2. \quad (10.2.29)$$

Note that $\hat{t} = 0$ only when $t = 0$. Then, the special solutions to the above equation satisfying the initial condition $a(0) = 0$ can be listed case by case as follows:

$$\text{case (a)} \quad (k = +1), \quad a = \frac{A}{2}(1 - \cos \hat{t}), \quad t = \frac{A}{2}(\hat{t} - \sin \hat{t}), \quad (10.2.30a)$$

$$\text{case (b)} \quad (k = 0), \quad a = \left(\frac{9A}{4}\right)^{1/3} t^{2/3}, \quad (10.2.30b)$$

$$\text{case (c)} \quad (k = -1), \quad a = \frac{A}{2}(\cosh \hat{t} - 1), \quad t = \frac{A}{2}(\sinh \hat{t} - \hat{t}). \quad (10.2.30c)$$

For each of these cases, the graph of $a(t)$ is similar to that in Fig. 10.9, so it is not shown here separately. The solution for the dust-only universe was first obtained by the Soviet physicist and mathematician Alexander Friedmann in 1922, and then independently by Georges Lemaître in 1927, much earlier than the discoveries by Howard P. Robertson and Arthur G. Walker in 1935. Therefore, the standard cosmological model is also often referred to as the **Friedmann-Lemaître-Robertson-Walker (FLRW) model**.

So far we have discussed the two extreme cases. The actual universe contains both matter and radiation. In this case, it is very difficult to solve the Friedmann equations quantitatively. However, a qualitative discussion is still possible. Firstly, observations indicate that our universe is presently in expansion, i.e., $\dot{a}(t_0) > 0$, with t_0 the present value of t . According to (10.2.18), \ddot{a} is negative, and hence the smaller $t > 0$ is, the greater $\dot{a}(t)$ is. Thus, for $0 < t < t_0$, the curve of $a(t)$ is convex upwards. Thus, the curve of $a(t)$ intersects the t -axis at a certain time before t_0 (which has been stipulated to be $t = 0$), similar to the curves in Fig. 10.9. Secondly, for $t > t_0$, we can write (10.2.20) as

$$\frac{d(\rho a^3)}{da} = -3pa^2. \quad (10.2.31)$$

Since p is always positive, the above equation indicates that ρa^3 decreases as a increases, and hence ρ decreases not slower than a^{-3} . Now rewrite (10.2.16) as

$$3(\dot{a}^2 + k) = 8\pi\rho a^2. \quad (10.2.32)$$

Then, as a increases, its right-hand side decreases not slower than a^{-1} . The above equation indicates that the behavior of \dot{a} depends on k . For $k = 0$, we have $\dot{a}^2 = \frac{8}{3}\pi\rho a^2$, which implies that \dot{a}^2 decreases as a increases, and that \dot{a} approaches zero as a goes to infinity. Noticing that $\dot{a}(t_0) > 0$, we see that $a(t)$ is positive for any $t > t_0$.

Hence, a increases as t increases, but the slope of the curve $a(t)$ keeps decreasing. Note that this does not ensure that a approaches infinity when t goes to infinity. In this case, the curve of $a(t)$ quantitatively behaves the same as the curve for $k = 0$ in Fig. 10.9. For $k = -1$, (10.2.32) results in $\dot{a}^2 = \frac{8}{3}\pi\rho a^2 + 1$, which is similar to the case of $k = 0$, only that the slope of the curve $a(t)$ tends to 1, instead of 0, as $a \rightarrow \infty$ (which corresponds to $t \rightarrow \infty$). This quantitatively behaves the same as the curve for $k = -1$ in Fig. 10.9. For $k = +1$, (10.2.32) results in $\dot{a}^2 = \frac{8}{3}\pi\rho a^2 - 1$, which indicates that \dot{a}^2 decreases as a increases. When $\frac{8}{3}\pi\rho a^2$ decreases to 1, \dot{a} decreases to zero. Let a_C be the value of a when $\frac{8}{3}\pi\rho a^2 = 1$, then it represents the critical state: in the process of a increasing from $a(t_0)$ to a_C , the value of \dot{a} decreases from $\dot{a}(t_0) > 0$, and becomes zero when a increases to be a_C (with the corresponding value of t denoted by t_C). Since \ddot{a} is always negative, according to (10.2.18), \dot{a} decreases as t increases. Consequently, \dot{a} becomes negative when $t > t_C$. That is, for $t > t_C$, a decreases as t increases until it becomes zero, and so a_C is obviously a maximum of a , which quantitatively behaves the same as the curve for $k = 1$ in Fig. 10.9. Thus, as we discussed case by case, the behavior of a for the actual universe can still roughly be described by Fig. 10.9.

[Optional Reading 10.2.1]

Rigorously speaking, it should also be proved that both values of a_C and t_C are finite. Let $f(a) \equiv a\dot{a}^2$. Then, due to $\dot{a}^2 = \frac{8}{3}\pi\rho a^2 - 1$, one has

$$f(a) = \frac{8}{3}\pi\rho a^3 - a = f_1(a) - f_2(a),$$

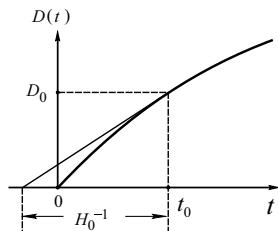
where $f_1(a) \equiv \frac{8}{3}\pi\rho a^3$ and $f_2(a) = a$. As shown in (10.2.31), the curve of f_1 is a curve with negative slope and positive function value. Thus, it is located in the first quadrant, and has an intersection with the graph of f_2 , a straight line with slope 1. It follows from $a\dot{a}^2 = f_1(a) - f_2(a)$ that the value of a at the intersection is exactly a_C , and thus a_C is finite. Let ρ_C be the value of ρ corresponding to a_C , then $\dot{a}^2 = \frac{8}{3}\pi\rho a^2 - 1$ results in $\rho_C = \frac{3}{8\pi a_C^2} > 0$. Since $p \geq 0$ in (10.2.18), we have $\ddot{a}(t_C) < 0$, or, more precisely, $\lim_{t \rightarrow t_C} \ddot{a}(t) < 0$, which guarantees that t_C is finite. In other words, it is impossible that a increases to a_C only when $t \rightarrow \infty$.

[The End of Optional Reading 10.2.1]

Since the universe has an initial time ($t = 0$), we can talk about its age in the present day, which is $t_0 - 0 = t_0$. Suppose $D(t)$ is the distance between two arbitrarily chosen galaxies, and denote $D(t_0)$ by D_0 for short. Roughly, assume that the universe expands at a constant rate, which is the presently observed rate u_0 . Then $t_0 \cong D_0/u_0 = D_0/H_0 D_0 = H_0^{-1}$. The value of H_0 is measured to be about 73 (km/s)/Mpc or 22.4 (km/s)/Mly.⁷ Hence, roughly, $t_0 \cong H_0^{-1} \cong 13.4$ Gyr. As we can see in Fig. 10.10, however, $t_0 < H_0^{-1}$ (see Exercise 10.4 for the precise relation of t_0 and H_0^{-1}), and thus t_0 is less than 13.4 billion years in the FLRW model. However, the observation for Type Ia supernovae in 1999 shows that the present universe is

⁷ Note that 1 Mpc = 10^6 pc, where pc stands for parsec (an abbreviation for “parallax second”), which is a commonly used unit of length in astronomy. Roughly speaking, 1 pc ≈ 3.26 ly (light-years). Also, 1 Mly = 10^6 ly and 1 Gly = 10^9 ly.

Fig. 10.10 The age of the universe $t_0 < H_0^{-1}$



not undergoing an decelerating expansion, as shown in Fig. 10.10, but an accelerating expansion, as shown in Fig. 10.13 (see Sect. 10.3.3 for details), which leads to $t_0 > H_0^{-1}$. This is a clear evidence that the FLRW model is not complete. The latest estimated value of the age of the universe is 13.787 ± 0.020 billion years [Aghanim et al. (2018)].⁸

For various reasons, it is difficult to measure an accurate value for H_0 . The value obtained by Hubble in the early days was about 8 times the value above, which gives a t_0 less than 1.7 billion years. This is unacceptable, because the age of the Earth is estimated to be 4.6 billion years based on the relative abundance of radioactive elements. Moreover, in the 1930s, the age of some stars had been incorrectly estimated to be 10^3 billion years, making it more unacceptable that the age of the universe is just 1.7 billion years. This made people raise questions regarding the theory of the Big Bang, and consequently many cosmological theories were brought out. At the end of the 1950s, when the value of H_0 was estimated to be about 1/8 of that measured by Hubble, those questions regarding the Big Bang theory were finally resolved.

For the study of cosmology, it is of great significance to improve the accuracy for the measurement of H_0 . For a long time, the maximum value of H_0 was as great as 2.5 times the minimum value. Usually it is regarded that

$$H_0 = 100h \text{ (km/s)/Mpc} \quad (10.2.33)$$

with $0.5 \lesssim h \lesssim 1$, which reflects the discrepancy between the estimated results. So far, the methods that have been used for measuring the Hubble constant include two major kinds. One is the “late universe” method, which measures the redshifts using the technique of a “calibrated distance ladder”. The values obtained from these measurements agree on a value near 73 (km/s)/Mpc [the latest data gives 73.30 ± 1.04 (km/s)/Mpc in Riess et al. (2022)]. The other is the “early universe” method, which is based on the CMB observations. The measurements of this kind have agreed on a value near 67.7 (km/s)/Mpc [67.66 ± 0.42 (km/s)/Mpc in Aghanim et al. (2018)]. Although the techniques of both kinds of measurements have been improved over the years and they both clearly converge on some certain values, these two values do not agree with each other. This discrepancy is called the **Hubble tension** [see Di Valentino et al. (2021) for a review].

⁸ The estimating and measurement of the age of the universe depends on the cosmological model, this result is based on the Λ CDM model (see Sect. 10.3.3).

As we have stated at the beginning of this chapter, the universe is the maximal spacetime containing everything in Nature. We should supplement this statement, noting that the universe described by the RW metric is just the universe after being “smoothed”, which only represents the behavior of the actual universe on a cosmological scale. If the local behavior of the actual universe in a relatively small scale is of concern, one should choose a suitable metric according to the local distribution of matter. Even though the universe is the maximal spacetime containing everything, the RW metric does not reflect the spacetime geometry of the local regions (i.e., those much smaller than cosmological scales).

10.2.4 The Cosmological Constant and Einstein’s Static Universe

As early in 1917, Einstein himself studied the universe using his field equation. Due to the widely accepted philosophical idea at that time that the universe is supposed to be invariant, he attempted to find a spacetime metric to describe a static universe. Unfortunately, the Einstein equation is not compatible with such a static solution. This is because static means $\dot{a} = 0$, and then (10.2.16) and (10.2.17) become

$$3k = 8\pi\rho a^2, \quad (10.2.16')$$

$$k = -8\pi p a^2, \quad (10.2.17')$$

which are obviously incompatible with the physical conditions $\rho > 0$ and $p > 0$. Einstein realized that there are no static solutions to his equation from the beginning. However, at that time he believed firmly that our universe is static, and so he modified his own equation just in order to acquire a static solution for the universe. He assumed that the modified field equation has the form $\tilde{G}_{ab} = 8\pi T_{ab}$. It follows from the properties of T_{ab} that \tilde{G}_{ab} must satisfy $\tilde{G}_{ab} = \tilde{G}_{ba}$ and $\nabla^a \tilde{G}_{ab} = 0$. For such a tensor field \tilde{G}_{ab} constructed out of g_{ab} and its derivatives of the first and second orders, \tilde{G}_{ab} can only be a linear combination of G_{ab} and g_{ab} (sans proof). Therefore, in 1917, Einstein published the modified Einstein equation

$$G_{ab} + \Lambda g_{ab} = 8\pi T_{ab}, \quad (10.2.34)$$

where Λ is a constant, called the **cosmological constant**.⁹

Now we will show that (10.2.34) indeed admits a static solution. First, we rewrite (10.2.34) as

$$G_{ab} = 8\pi (T_{ab} - \Lambda g_{ab}/8\pi), \quad (10.2.35)$$

⁹ Einstein assumed that Λ is very small, so that the Λ -term is negligible in every other problem except for cosmology [see Rindler (1982)].

and formally treat $T_{ab} - \Lambda g_{ab}/8\pi$ as a new “energy-momentum tensor”. In this manner, (10.2.35) is still, formally, Einstein’s equation without the cosmological constant. For convenience, the actual energy-momentum tensor T_{ab} will be now denoted by \bar{T}_{ab} , and T_{ab} will denote the new energy-momentum tensor, i.e., $\bar{T}_{ab} - \Lambda g_{ab}/8\pi$. Then, (10.2.35) can now be expressed as

$$G_{ab} = 8\pi T_{ab} = 8\pi (\bar{T}_{ab} - \Lambda g_{ab}/8\pi), \quad (10.2.36)$$

In the original model, there is only matter (dust), but no radiation, i.e., \bar{T}_{ab} depends only on $\bar{\rho}$ but not \bar{p} . Then, $\bar{T}_{ab} = \bar{\rho} U_a U_b$, and thus $\bar{T}_{00} = \bar{\rho}$, $\bar{T}_{ij} = 0$. It follows from (10.2.13) that the ρ and p in the new energy-momentum tensor T_{ab} satisfy

$$\rho = T_{00} = \bar{T}_{00} - \Lambda g_{00}/8\pi = \bar{\rho} + \Lambda/8\pi, \quad (10.2.37)$$

$$pg_{ij} = T_{ij} = \bar{T}_{ij} - \Lambda g_{ij}/8\pi = -\Lambda g_{ij}/8\pi, \quad (10.2.38)$$

where (10.2.38) is also equivalent to

$$p = -\frac{\Lambda}{8\pi}. \quad (10.2.38')$$

This indicates that the introduction of the Λ -term is equivalent to adding “matter” with a negative pressure p into the universe (as long as $\Lambda > 0$). In this case, the equation system of (10.2.16') and (10.2.17') will admit a solution. Plugging (10.2.38') into (10.2.17) yields

$$k = a^2 \Lambda. \quad (10.2.39)$$

On the other hand, plugging (10.2.37) into (10.2.16') yields

$$3k = (8\pi \bar{\rho} + \Lambda)a^2. \quad (10.2.40)$$

Subtracting the two equations above, we have

$$2k = 8\pi \bar{\rho}a^2. \quad (10.2.41)$$

The condition $\bar{\rho} > 0$ leads to $k > 0$, and hence

$$k = +1. \quad (10.2.42a)$$

Then, (10.2.39) and (10.2.41) give, respectively,

$$\Lambda = a^{-2}, \quad (10.2.42b)$$

$$a^2 = \frac{1}{4\pi \bar{\rho}}. \quad (10.2.42c)$$

Equation (10.2.42) represents the unique static solution for a dust-only universe with the Λ -term added, where $\bar{\rho}$ is the density of the dust. Equation (10.2.42a) indicates that the spatial geometry of this solution is spherical, with the corresponding 4-dimensional line element

$$ds^2 = -dt^2 + a^2 [d\psi^2 + \sin^2 \psi (d\theta^2 + \sin^2 \theta d\varphi^2)], \quad (10.2.43)$$

called the metric of **Einstein's static universe**, where $a^2 = 1/\Lambda$ is a constant. Although (10.2.43) describes a static solution, this is not a stable solution, which will turn into a contracting or an expanding solution once we apply a perturbation to it. However, since Einstein did not write down the differential equations for $a(t)$, he did not notice the instability, nor did anyone else until A. S. Eddington in 1930 [see Ellis (1989)].

Because Einstein firmly believed that the universe is static, he refuted, as a referee, the paper submitted to *Zeitschrift für Physik* by Friedmann in 1922, and neglected Friedmann's appealing letter to him. In 1923, Einstein felt that Friedmann's paper was plausible, after Yuri Krutkov, a friend of Friedmann, explained Friedmann's opinions to him. Then, Einstein sent the retraction of his refutation to the journal, saying that he was convinced that Friedmann's results are “correct and shed new light”. Einstein abandoned the cosmological constant in 1931, after Hubble's confirmation that the universe is expanding. It is reported that Einstein referred to the introduction of the cosmological constant as his “biggest blunder” [Peebles and Ratra (2003)]. However, the story of the cosmological constant does not end here. Over the past century, its profound impact on cosmology, and even the entirety of physics, has developed in many startling ways. We will get back to this in Sect. 10.3.3 and Chap. 15 in Volume II.

10.3 The Thermal History of Our Universe

10.3.1 A Brief History of the Universe

As we have discussed, the universe is radiation-dominated when the scale factor a is sufficiently small. The transition to the matter-dominated universe happens at about $t = 10^{11}$ s, i.e., thousands of years after the big bang. Based on the standard cosmological model, in this subsection we will briefly introduce the history of our universe's evolution up to today. Some basics of high energy physics will be inevitably involved in the discussion, so it will be helpful if the reader has learned some basic knowledge before. However, to avoid too much particle physics, thermodynamics and quantum statistical mechanics, we will only provide a rough introduction [for more precise, more detailed discussions, see Kolb and Turner (1990); Weinberg (2008)]. Since the universe contains everything in Nature, and does not have an exterior, its evolution can be regarded as an adiabatic expansion. The early universe is radiation-

dominated, and it is not difficult to derive the relation between the temperature T and the scale factor a . It follows from (10.2.22) that the energy density ρ of the radiation-dominated universe is proportional to a^{-4} , and from quantum statistical mechanics we know that ρ is proportional to T^4 for radiation,¹⁰ and hence $T \propto a^{-1}$. On the other hand, the k in (10.2.24) is negligible when a is sufficiently small, and thus its solution is $a = (2Bt)^{1/2}$, combined with $T \propto a^{-1}$ yields $Tt^{1/2} = \text{constant}$. The value of this constant in SI is about 10^{10} , and hence

$$T = \frac{10^{10}}{\sqrt{t}} \quad (\text{the units for } T \text{ and } t \text{ are K and s, respectively}). \quad (10.3.1)$$

This is an approximate relation of the temperature T and time t for the early universe (radiation-dominated), from which we can see that $T = \infty$ at $t = 0$ (the big bang). Therefore, starting from the big bang singularity where the temperature is infinitely high, the evolution of our universe is a process of adiabatic expansion with the temperature continually decreasing.

1. The big bang singularity.

The expansion of the universe starts from the big bang singularity ($t = 0, T = \infty$). The spacetime singularity is one of the thorniest problems. Many physical quantities approach infinity as one approaches the singularity, where all the physical laws also become invalid. Before 1965, most of the physicists did not believe in the existence of a spacetime singularity, and tried to post various reasons for avoiding singularities. Making use of global differential geometry, R. Penrose and S. W. Hawking proved, first individually and then jointly, a series of singularity theorems, which assert that spacetime singularities (including the collapse of a star in its late stage and the big bang at the beginning of the universe) are inevitable as long as some reasonable conditions are satisfied (see Appendix E in Volume II for a qualitative introduction to singularity theorems). What is notable is that these conditions do not contain any requirement on symmetry. Subsequently, many relativists had to admit the existence of singularities, and so a variety of intensive studies regarding singularities sprung up. However, since it is hard to believe that physical quantities can be infinite, one may look at singularity theorems from another perspective: rather than prove the existence of singularities, singularity theorems indicate classical general relativity fails to be applicable near a singularity (where the spacetime curvature is very large). As is well-known, there were two great revolutions of physics that happened in the early 20th century—the creation of relativity and quantum theory. In the perspective of understanding the spacetime structure and the essence of gravity, general relativity is undoubtedly a revolutionary theory, while it is “not quite revolutionary” from another perspective, since it does not obey the fundamental principles of quantum theory. According to quantum theory, any observable cannot have a determined value (unless

¹⁰ For electromagnetic radiation, it follows from the law of blackbody radiation that $\rho \propto T^4$. If one considers the contributions from other particles, ρ and T will have the relation $\rho = (\pi^2/30)N_{\text{eff}}T^4$, where N_{eff} is a number determined by the number of types of the particles whose rest energy is far less than $k_B T$ (k_B is the Boltzmann constant). Thus, $\rho \propto T^4$ only when N_{eff} is a constant.

the system is in an eigenstate of this observable), and one can only make probabilistic predictions for the results of a measurement. However, all the observables (e.g., the metric) in general relativity have determined values (as we describe the history of a particle using its world line, we have assumed that it has a determined position at each moment). Nowadays, it has been a consensus that a theory which does not consider quantum effects is referred to as a classical theory, and thus Einstein's general relativity is referred to as classical general relativity.¹¹ Since singularity theorems indicate that classical general relativity breaks down when the spacetime curvature is sufficiently large, there should exist a critical time $t_C > 0$ in the very early universe, where classical general relativity is invalid in the period $[0, t_C]$ and should be substituted by a brand new theory of **quantum gravity**. Although people have been exploring for this quantum theory of gravity actively and important progress keeps being made, so far we have not established a complete theory yet. Hence, we still cannot consider the singularity or a region very close to it (within $[0, t_C]$) and our discussion can only start from the critical time t_C . How do we estimate the value of t_C ? Since this question involves spacetime, gravity and quantum theory, t_C should only depend on fundamental constants c , G and \hbar , and the “unique” quantity with time dimension constructed by c , G and \hbar is the **Planck time** $t_P \equiv (G\hbar/c^5)^{1/2} \sim 10^{-43}$ s. Therefore, t_P is taken as the critical time t_C , i.e., a rough bound for the region where classical general relativity is valid is at t_P (see Optional Reading 10.3.1 for details). We will only discuss the history of the evolution after $t_P \sim 10^{-43}$ s.

[Optional Reading 10.3.1]

Is can be said that the spacetime curvature in the period $[0, t_C]$ is so large that classical general relativity breaks down. However, this statement needs some explaining. First of all, what is the magnitude of the spacetime curvature? The spacetime curvature is a tensor, whose magnitude usually refers to a scalar constructed from the curvature tensors (and metric), such as the scalar curvature $R \equiv g^{ab}R_{ab}$ and the scalar $\mathcal{R} \equiv R^{ab}R_{ab}$. The early universe is radiation-dominated, and the trace of the energy-momentum tensor of the electromagnetic radiation (a null electromagnetic field) gives $T = 0$, and so from Einstein's field equation we can see that $R = 0$ in this case. Therefore, we may use $\mathcal{R} \equiv R^{ab}R_{ab}$ to represent the magnitude of the spacetime curvature of the early universe. Second, what value of \mathcal{R} is large enough so that classical general relativity is invalid? We would like to find a critical value \mathcal{R}_C such that in a very rough sense we can say that classical general relativity is valid when $\mathcal{R} < \mathcal{R}_C$ and it is not when $\mathcal{R} > \mathcal{R}_C$. The most solid way is to determine this bound by a theory of quantum gravity, but we do not have such a theory yet. A concession would be to obtain some information using perturbation techniques, from which one can get an approximate order of magnitude of \mathcal{R}_C . Another cursory but quite convenient method is dimensional analysis. The dimension of \mathcal{R} in SI is L^{-4} [which can be derived from (A.7) in Appendix A], while the “unique” quantity with length dimension constructed by c , G and

¹¹ Note that the criterion for “classical physics” has become different from that in the first half of the 20th century. People used to refer to (both special and general) relativity and quantum mechanics as “modern physics” and the previous physics as “classical physics”. As time goes on (especially as people realized that general relativity has to be combined with quantum theory), the term “classical” gradually becomes a synonym of “non-quantum”, and the general relativity without considering quantum effects is referred to as classical general relativity to be distinguished from a theory of quantum gravity. As this criterion for “classical physics” has become a consensus among physicists internationally, the previous interpretation of the word “classical” now seems to be too “classical”.

\hbar is the **Planck length** $l_P \equiv (G\hbar/c^3)^{1/2} \sim 10^{-35}$ m. Hence, it is generally accepted that $\mathcal{R}_C \sim l_P^{-4}$ (\sim means they are of the same order).

In a word, one can roughly regard $\mathcal{R}_C \sim l_P^{-4}$ by means of dimensional analysis. On the other hand from dimensional analysis we also have $t_C \sim t_P$. It is natural to ask: if we assume for now that classical general relativity is applicable, would the value of \mathcal{R} really have the same magnitude as $\mathcal{R}_C \sim l_P^{-4}$ when the universe evolves to t_P ? From the expressions of the Christoffel symbols below (10.2.5) we can find all the nonvanishing components of the Ricci tensor of the FLRW universe as

$$\begin{aligned} R_{00} &= -3\ddot{a}/a, & R_{11} &= (1 - kr^2)^{-1}(a\ddot{a} + 2\dot{a}^2 + 2k), \\ R_{22} &= r^2(a\ddot{a} + 2\dot{a}^2 + 2k), & R_{33} &= r^2 \sin^2 \theta(a\ddot{a} + 2\dot{a}^2 + 2k), \end{aligned}$$

from which we obtain

$$\mathcal{R} \equiv R^{\mu\nu} R_{\mu\nu} = 9(\ddot{a}/a)^2 + 3a^{-4}(a\ddot{a} + 2\dot{a}^2 + 2k)^2. \quad (10.3.2)$$

Since the very early universe is radiation-dominated, and since we can take $k = 0$ when a is very small, it follows from (10.2.25b) that $a(t) = (2Bt)^{1/2}$. Taking the derivative we find $-\ddot{a}/a = (\dot{a}/a)^2 = (1/4)r^{-2}$, then plugging in (10.3.2) yields $\mathcal{R} = 0.75t^{-4}$. Converting back to SI (adding c to it), one finds that the value of \mathcal{R} at t reads $\mathcal{R}(t) = 0.75c^{-4}t^{-4}$. Noticing that $l_P = c t_P$, we have

$$\mathcal{R}(t_P) = 0.75l_P^{-4} \sim l_P^{-4}.$$

That is, the curvature $\mathcal{R}(t_P)$ at t_P is of the same order as $\mathcal{R}_C \sim l_P^{-4}$, and thus classical general relativity is not applicable in the period $[0, 10^{-43}$ s].

[The End of Optional Reading 10.3.1]

2. Thermal equilibrium in the early universe.

Although classical general relativity begins to be valid since about $t = 10^{-43}$ s, the temperature is extremely high in a small period of time right after $t = 10^{-43}$ s, and such a high energy is still too high to tackle for high energy physics. According to the Standard Model of particle physics, the universe consists of particles and antiparticles with extremely high energy, including quarks, leptons, gauge bosons (e.g., photons) that mediate interactions, and the Higgs boson which is associated with a mechanism that gives masses to other elementary particles.¹² The frequent interactions between these high energy particles make them live in thermal equilibrium, which may be described as “a pot of thoroughly stirred soup of elementary particles” (the “pot” refers to the whole space of the universe at a certain time). Take a photon γ as an example, it could not travel for a long path as unimpeded as it could in the present universe; its mean free time (i.e., the average time between collisions) is very short due to the frequent collisions with other particles (including scattering, absorption and emission). Although the universe is also expanding rapidly, the photon would have collided with other particles numerous times before it “notices” that the universe has expanded. Besides photons, neutrinos have the same experience in the very early universe. In a word, although the universe keeps expanding, the rate of the interactions between particles is greater than the rate of the expansion of the universe during most of the universe’s history (especially the early universe). To put it intuitively, the speed

¹² Besides, there might also be other particles beyond the Standard Model that have not been discovered yet.

of “stirring” is way faster than the speed of the expansion of the “pot” (the whole space). Therefore, in most parts of the early universe all kinds of particles can reach a local thermal equilibrium.

According to quantum statistical physics, the average energy of the radiation particles emitted in the radiation with a temperature T is roughly equal to $k_B T$, where k_B is the Boltzmann constant. This conclusion can also be approximately applied to matter particles with rest energy far less than $k_B T$, whose speed is close to the speed of light. Together with radiation particles, they are called **relativistic particles**. For example, $k_B T \cong 10$ MeV when $T = 10^{11}$ K, while the rest energy of an electron e is about 0.5 MeV, and hence an electron is a relativistic particle when $T = 10^{11}$ K. According to quantum field theory, two photons can be transformed into some particle-antiparticle pair (“pair production”), and a particle-antiparticle pair can also be transformed into two photons (“pair annihilation”). Of course, both of these two processes satisfy the energy conservation law. The average energy $k_B T$ of a photon at room temperature is far less than the rest energy of an electron, and thus the probability of two photons becoming an electron-positron pair ($2\gamma \rightarrow e + e^+$) is almost zero. However, at a high temperature like $T = 10^{11}$ K, the rate of this kind of “pair production” is very large (basically proportional to the density of photons). When e and e^+ collide, they can also annihilate into two photons ($e + e^+ \rightarrow 2\gamma$), and the rate of the annihilation is proportional to the density of (e, e^+) pairs. Therefore, when equilibrium is reached, the density of (e, e^+) pairs is roughly equal to the photon pairs whose energy is greater than the rest energy m_e of an electron. Conversely, since the rest energy of a proton p and a neutron n is about 1840 times the rest energy m_e of an electron, the densities of (p, \bar{p}) and (n, \bar{n}) pairs are almost zero even at this high temperature $T = 10^{11}$ K (where \bar{p} and \bar{n} stand for antiproton and antineutron, respectively).

3. Asymmetry of matter and antimatter.

When $t = 1$ s and $T = 10^{10}$ K, because $k_B T \gg m_e$ and $k_B T \ll m_p$, there exist plenty of (e, e^+) (the same order as the number of γ) while there are almost no (p, \bar{p}) and (n, \bar{n}) . Therefore, the contents of the universe are: a large amount of neutrinos ν and antineutrinos $\bar{\nu}$, a large amount of photons γ , a large amount of (e, e^+) (the number density of each kind of particle above is basically the same) and a small amount of protons p and neutrons n . Earlier than this, such as when $T \gg 10^{13}$ K, since $k_B T > m_p$, there used to be a large amount of (p, \bar{p}) and (n, \bar{n}) , which vanished due to annihilation when the temperature decreased to $k_B T < m_p$. Since the p, \bar{p} and n, \bar{n} annihilate in pairs, why could there still be a small amount of p and n that remain? The reason we know that there must be a small amount of p and n is because the matter in the present universe are all composed of p and n , while antiparticles in the present universe are extremely rare. That is to say, there exists a particle-antiparticle (matter-antimatter) asymmetry in the present universe. If we accept this fact, we have to admit that besides a large amount of (p, \bar{p}) and (n, \bar{n}) , there should be a small amount of unpaired p and n before $t = 0.01$ s. When $k_B T < m_p$, p and \bar{p} , n and \bar{n} annihilate in pairs, with only a small amount of p and n (both are baryons) remaining.

It is estimated that n_b/n_γ , the ratio of the number densities of baryons and photons in the universe, is only on the order 10^{-10} , but it is surely not zero.

If one questions further about the source of this asymmetry of baryons and antibaryons, then there are only two possible answers: ① The universe prefers particles over antiparticles from its beginning (which is obviously not quite natural); ② There were the same number of baryons and antibaryons at the beginning of the universe, and for some reason baryons became favored during its very early evolution. If one believe that the baryon number must be conserved, then the latter choice would not be acceptable. Fortunately, there could be ways to bypass this difficulty. For example, a Grand Unification Theory (GUT) proposed in the 1970s which unifies the electromagnetic, weak, and strong interactions suggests that the baryon number may not be conserved at a very high energy scale. It has been shown that the baryon number not being conserved plus a temporary deviation from thermal equilibrium in the very early universe may create surplus p and n from the universe which originally has particle-antiparticle symmetry. Although the present GUT models have not been supported by experiments as anticipated, it is generally believed that a successful Grand Unification Theory will sooner or later resolve the above difficulty in cosmology.

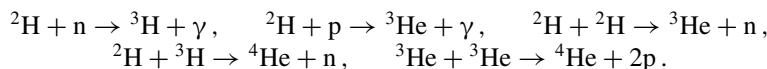
4. Neutrino decoupling.

When $t = 1$ s and $T = 10^{10}$ K, $k_B T \cong 1$ MeV is still greater than m_e , and hence there still exist plenty of (e, e^+) . However, since the temperature and density have decreased a lot compared with before, the interaction rate between neutrinos (or antineutrinos) and other particles is far less than the expansion rate of the universe. The mean free time of a neutrino got extended significantly, and so it becomes approximately a free particle that does not interact with other particles, which means it is no longer in thermal equilibrium with other particles. This is called the **decoupling** of neutrinos. The decoupling time and temperature of neutrinos are denoted by t_{vd} and T_{vd} , respectively. Although neutrinos will fill up the universe after the decoupling just like other particles, and keep affecting the evolution of the universe since they still contribute to the total energy-momentum tensor, they are not correlated with any other constituents of the universe in any other aspects. This huge amount of neutrinos evolve independently up to today, and they now exist as an independent particle system whose temperature is about 1.95 K, known as the cosmic neutrino background. Since the interaction between neutrinos and a detector is extremely small, it is almost impossible to observe the cosmic neutrino background directly. However, indirect evidence has been observed from the fluctuations of the cosmic microwave background [see Follin et al. (2015)].

5. Primordial nucleosynthesis.

Observations indicates that about 1/4 of the total baryonic mass of the current universe is helium. Except for primordial nucleosynthesis in the early universe, there is no other known process that could have created this abundance of helium. (Although the nuclear reactions inside stars keep producing helium, they only contribute to a small portion of the abundance above.) The temperature involved in primordial nucleosynthesis is roughly between 10^{10} K and (slightly lower than) 10^9 K. When

the temperature is higher than this range, even if a proton and a neutron could combine into a helium nucleus, it will be shattered by the high energy photons (called “photofission”). Since the physics in this temperature interval is already well-studied and has been confirmed in labs on the Earth, people are confident enough for the theory of primordial nucleosynthesis. Due to the fact that the density number of the nuclei is relatively low and the rapid expansion of the universe leads to a very short reaction time (about 10^2 s), a reaction can only happen for two high speed particles in primordial nucleosynthesis. First, protons and neutrons combine into deuterons, and the rest of the energy and momentum is carried away by a photon ($p + n \rightarrow {}^2H + \gamma$). And then this is followed by a sequence of reactions which produce 3H (triton), 3He and 4He , such as



Since there does not exist a stable nuclide whose mass number is 5, the reaction chain stops here. As the main product, 4He gradually accumulates, and the nuclear reaction continues until there are a large enough number of nuclei, which will lead to the production of a tiny amount of 7Li . Since there does not exist a stable nuclide whose mass number is 8, the reaction chain again ends here. The first step that this whole reaction chain must go through is protons and neutrons combining into deuterons. The binding energy of a deuteron nucleus is much lower than that of a helium nucleus. A helium nucleus can remain stable when the temperature decreases to 3×10^9 K, while at this temperature a deuteron nucleus will be broken right after it is formed. Therefore, the nucleosynthesis process that is actually meaningful begins after the “deuteron barrier” is passed when the temperature is slightly lower than 10^9 K, and the product is a large amount of 4He and a tiny amount of 2H , 3He and 7Li (3H is unstable and will quickly decay to 3He). If we take the yield of 4He as the unit, then the yields of 2H and 3He are about 10^{-5} , while the yield of 7Li is about 10^{-10} . As for all kinds of elements that are heavier than 7Li in today’s universe, they mainly come from the nuclear reactions in the interior of stars and supernova explosions. The reason that the reactions inside a star can skip over the elements with $A = 5$ and $A = 8$ and yield heavy elements is that the self-gravity there is so strong that the density of the star’s core is extremely high, and there is enough reaction time such that three-particle collisions can happen.

The helium abundance produced by the primordial nucleosynthesis closely depends on the ratio n_n/n_p of the number density of protons and neutrons before the end of the nucleosynthesis (the reason will be seen shortly). This ratio can be derived from the following discussion. Before the neutrinos decouple, protons and neutrons can convert mutually by the following weak interaction processes: $p + e \leftrightarrow n + \nu_e$, $p + \nu_e \leftrightarrow n + e^+$. Since the mass of a neutron is slightly greater than the mass of a proton ($m_n - m_p \cong 2.5m_e$), it is more difficult for a proton to turn into a neutron than the reverse. For example, since $m_p + m_e \cong m_n - 1.5m_e < m_n$, it follows from the conservation of energy that a rest proton and a rest electron cannot even turn into a rest neutron, but the reverse process does not have such an issue. Certainly, the

energy of an electron when the temperature is above 10^{10} K is far greater than its rest energy m_e , and thus $p + e \rightarrow n + \bar{v}_e$ could happen, but nevertheless its probability is always less than that of the reverse process. Therefore, n_n/n_p should be less than 1 when the forward and reverse reactions reach a statistical equilibrium, and the quantitative relation is given by the Boltzmann equation

$$\frac{n_n}{n_p} = e^{-\frac{\Delta m}{k_B T}}, \quad (10.3.3)$$

where $\Delta m \equiv m_n - m_p$. When the temperature drops to $T_{vd} \cong 10^{10}$ K, the neutrinos are decoupled and the above weak interaction processes of n and p converting into each other basically stop, then n_n/n_p will almost freeze out at the value $e^{-\frac{\Delta m}{k_B T}}$. The discussion above neglected the spontaneous decay of free neutrons, $n \rightarrow p + e + \bar{v}_e$, since its half-life (about 10 mins) is far greater than the age of the universe at T_{vd} ($t_{vd} \cong 1$ s). However, the age of the universe when helium is synthesized ($t \cong 10^2$ s) already takes a considerable portion of the half-life of neutrons, and thus n_n/n_p will be slightly lower than its freeze-out value $e^{-\frac{\Delta m}{k_B T_{vd}}}$, which is a little below 1/7. Let N_n and N_p represent the total neutron number and total proton number, and let $\sigma \equiv N_n/N_p$, then the total nucleon number $N = N_n + N_p = (\sigma + 1)N_p$, where all the neutrons are combined with the same number of protons and turn into helium, and hence the nucleon number contained in helium is $N_{He} = 2N_n = 2\sigma N_p$. Therefore, the helium abundance (measured by mass) produced by the primordial nucleosynthesis is

$$Y = \frac{N_{He}}{N} = \frac{2\sigma N_p}{(\sigma + 1)N_p} = \frac{2\sigma}{\sigma + 1},$$

i.e.,

$$Y = 2 \left(\frac{n_n}{n_p} \right) \left(1 + \frac{n_n}{n_p} \right)^{-1}. \quad (10.3.4)$$

Plugging in $n_n/n_p \cong 1/7$ yields $Y \cong 0.25$. Apart from primordial nucleosynthesis, the nuclear reactions in the interior of stars also produces ${}^4\text{He}$ (a lot less than the production of primordial nucleosynthesis though), and thus it is necessary to deduce the primordial helium abundance (the abundance of helium when primordial nucleosynthesis is over) from the observed helium abundance. As the accuracy of measurements got improved over the years, recent estimations [$Y = 0.245 \pm 0.003$ in Zyla et al. (2020)] have matched very well with the theoretical value above. Although the abundances of other products (${}^2\text{H}$, ${}^3\text{He}$ and ${}^7\text{Li}$) are very small, they are also significant for verifying the theory. There is another important physical parameter η involved in the quantitative calculation of the abundances of the products of primordial nucleosynthesis, which is defined as the ratio of the densities of the baryons and photons in the universe ($\eta \equiv n_b/n_\gamma$). η^{-1} stands for the photon number around each baryon, which affects the starting time of the nucleosynthesis by affecting the difficulty of photofission, and thus affects the abundances of the products. The abun-

dance of ${}^4\text{He}$ only depends weakly on η , while the abundances of ${}^2\text{H}$, ${}^3\text{He}$ and ${}^7\text{Li}$ are rather sensitive to η . Calculation shows that as long as one assumes that η is in the range of $5.8 \times 10^{-10} \sim 6.5 \times 10^{-10}$, i.e.,

$$\eta \equiv \frac{n_b}{n_\gamma} \cong (5.8 \sim 6.5) \times 10^{-10}, \quad (10.3.5)$$

then the theoretical abundances of all four products above agree with their observational abundances [Zyla et al. (2020)]. This not only is a powerful support to the theory of nucleosynthesis, but also sets a rather clear (and narrow) possible range for this key parameter η , which provides another important contribution to cosmology.

Another important contribution of the theory of primordial nucleosynthesis is that it determines the number of neutrino species N_ν as 3, i.e., it confirms that there are only 3 types of neutrinos (and thus leptons only have three generations). This is supposed to be a problem of particle physics; the history of cosmology being used in this subject started from 1976. The situation of high energy physics at that time had the following features: ① there was already evidence that, beside the first two generations of leptons e and μ (and their corresponding neutrinos ν_e and ν_μ), there exists a third generation of leptons (and thus a third type of neutrino); ② the accelerators at that time could not provide any meaningful restriction on N_ν ; ③ many physicists tended to believe that the value of N_ν would increase as the energy of the accelerators increased; ④ very few particle physicists believed that the study of cosmology could be helpful to particle physics. However, G. Steigman and collaborators blazed a new trail by pointing out that the increase of the number of neutrino species will lead to an increase in the abundance of ${}^4\text{He}$ coming from primordial nucleosynthesis, and thus the observed abundance of ${}^4\text{He}$ should give an upper bound for N_ν . The basic idea is as follows: since the k in (10.2.16) is negligible when a is small, it is easy to see from $H \equiv \dot{a}/a$ that $H^2 = 8\pi\rho/3$. More species of neutrinos leads to a greater ρ , which leads to a greater H due to the equation above, namely a faster expansion of the universe. This would make neutrinos decouple earlier, i.e., t_{vd} would be smaller, and thus the decoupling temperature T_{vd} would be greater. It follows from (10.3.3) that this “freezes out” n_n/n_p at a greater value, and hence the abundance of ${}^4\text{He}$ would be higher. The upper bound they gave in Steigman (1977) was $N_\nu \leq 7$. This article demonstrated the novel insight that “cosmology can provide important constraints on particle physics, and the universe is an important supplement for high energy accelerators.” Later on, more and more studies have been carried out along this direction, which keeps shrinking down the estimated value of N_ν [see Steigman (2012) for a review]. A recent analysis in Cyburt et al. (2016) gives $N_\nu \leq 3.2$, which agrees well the result $N_\nu = 3$ obtained from the collider experiments by the European Organization for Nuclear Research (CERN).

6. Cosmic microwave background radiation.

In a long period of time after primordial nucleosynthesis, nothing significant happens in the universe until $t \cong 10^{13}$ s $\cong 4 \times 10^5$ years at which time $T \cong 3000$ K (or 4000 K). At this temperature, nuclei and electrons start to combine into neutral atoms

(before this the electrons still have enough energy to escape from the electromagnetic bound of a nucleus), and the matter in the universe starts to transfer quickly from an ionized state (plasma) to the neutral state. In an ionized state, photons interact frequently with charged particles (especially free electrons), and thus they are in thermal equilibrium with the matter particles. However, photons have almost no interaction with neutral particles, and thus the universe becomes transparent after the charged particles are combined into neutral particles (the mean free time of a photon is a lot longer than the present age of the universe). At this stage, photons are decoupled from the “big family” of the particles in thermal equilibrium and become an independent system. Before decoupling, these photons were in thermal equilibrium with the matter particles (similar to the photons in an oven being in thermal equilibrium with the particles of the oven’s wall), whose energy density distribution in wavelength satisfies the blackbody radiation curve, which can be described by Planck’s law:

$$du = \frac{8\pi hc}{\lambda^5} \left(e^{\frac{hc}{k_B T \lambda}} - 1 \right)^{-1} d\lambda, \quad (10.3.6)$$

where du stands for the energy per unit volume of the photons whose wavelength is in the range $(\lambda, \lambda + d\lambda)$, T is the temperature, and h and k_B are the Planck constant and Boltzmann constant, respectively. Although the photons are no longer in thermal equilibrium with the matter particles after decoupling, their energy distribution in wavelength still satisfies Planck’s law, only the temperature T will decrease inversely as the scale factor a increases. The reason can be briefly explained as follows: after the photon decoupling, suppose a is increased by a factor α , i.e., $a' = \alpha a$, then the number of photons per unit volume is decreased by a factor α^{-3} . On the other hand, the energy of each photon also decreases by a factor α^{-1} due to the redshift [see (10.2.8)]. Therefore, the energy of those photons whose wavelength is in the range $(\lambda, \lambda + d\lambda)$ per unit volume (when the scale factor is a) will decrease to

$$du' = \alpha^{-4} du = \frac{8\pi hc}{\alpha^4 \lambda^5} \left(e^{\frac{hc}{k_B T' \lambda}} - 1 \right)^{-1} d\lambda,$$

Expressed in terms of the new wavelength $\lambda' = \alpha \lambda$, this becomes

$$du' = \frac{8\pi hc}{\lambda'^5} \left(e^{\frac{hc}{k_B T' \lambda'}} - 1 \right)^{-1} d\lambda', \quad \text{where } T' \equiv \alpha^{-1} T. \quad (10.3.7)$$

Thus, the distribution of energy density in wavelength when the scale factor increases to α' can still be described by Planck’s law, which just corresponds to a lower temperature T' . Estimation shows that the temperature of the decoupled photon system in the present day is $T_0 \sim 3$ K. That is to say, the present universe is filled with a large amount of background photons homogeneously (all the galaxies are “soaked” in the bath of ubiquitous photons), and the distribution of their energy in wavelength is described by the blackbody radiation curve at 3 K. The radiation energy is mainly concentrated in microwave band (the wavelength of the maximum energy density

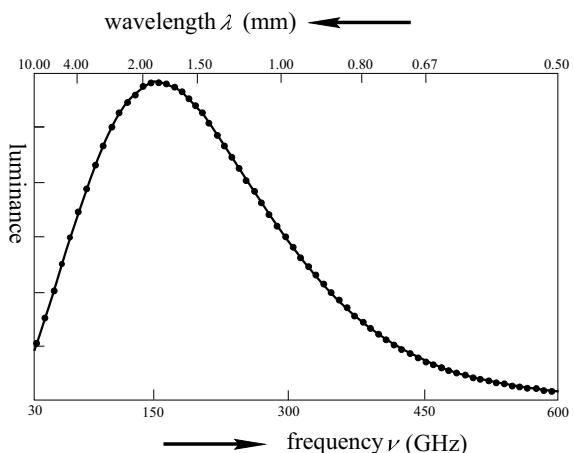
is about 0.1 cm), and therefore this is called the **cosmic microwave background radiation (CMB, CMBR)**.

American physicists and radio engineers A. A. Penzias and R. W. Wilson detected this isotropic radiation accidentally in 1965, and received the 1978 Nobel Prize in Physics for this discovery. What they detected was in fact the signal at only one wavelength (7.35 cm) (i.e., only one point on the curve). Assume that this is blackbody radiation, then the temperature corresponds to the blackbody radiation curve passing through this point is 3.5 K. American physicist R. H. Dicke and colleagues pointed out immediately that this is a trace (a “fossil”) of the big bang, which was exactly the cosmic background radiation they were preparing to search for. However, there were also a few articles at that time which gave alternative explanations for this signal. In order to confirm that this is indeed a trace of the big bang, two conditions need to be satisfied: ① the distribution of the energy spectrum is a blackbody radiation curve; ② the radiation is highly isotropic [the intensity (or the corresponding temperature) is the same in all directions]. This urged people to measure the other points of the curve and to test the isotropy of the radiation. Soon (in 1967), it was confirmed that the anisotropy is no more than 0.1–0.3%, and the results of measuring many other points with wavelength greater than 0.3 cm all fit the blackbody radiation curve. The radiation with wavelength less than 0.3 cm can be easily absorbed by the atmosphere, which could be measured outside the atmosphere by balloons or satellites. Since 1989, the Cosmic Background Explorer (COBE) satellite started to measure for a wide wave band in a high precision and obtained a perfect blackbody radiation curve. Figure 10.11¹³ illustrates the first results published in 1990, which is regarded as the most perfect blackbody radiation observed by humans in nature.

COBE also presented a more precise result for the anisotropy of the background radiation. Expand the temperature T (as a function of the angular coordinates) in terms of the spherical harmonics, then other than the constant term T_0 , the two lowest order spherical harmonics are called the **dipole moment** and **quadrupole moment**, which are the main manifestations of the anisotropy. Let T_1 and T_2 represent the amplitudes of the dipole anisotropy and quadrupole anisotropy, respectively, then the measurements of COBE give $T_1/T_0 \sim 10^{-3}$ and $T_2/T_0 \sim 10^{-5}$. The former can be reasonably interpreted as the consequence of the small velocity of the Earth relative to the isotropic reference frame: the Earth orbits around the Sun, the Sun moves relative to the center of the Milky Way, and the Milky Way also has “peculiar motion” relative to the isotropic reference frame. By definition, only isotropic observers can obtain isotropic results from measurements, and thus it certainly makes sense that a small anisotropy of the background radiation is observed by the Earth’s observer. Analysis shows that the first-order approximation of this anisotropy manifests exactly as the

¹³ The luminance B_ν in this figure refers to the energy per unit frequency transmitted per unit area, per unit solid angle, per unit time, whose unit is $\text{J}\cdot\text{s}^{-1}\cdot\text{m}^{-2}\cdot\text{sr}^{-1}\cdot\text{Hz}^{-1}$, where sr stands for steradian. The corresponding luminance of u in (10.3.6) is not B_ν but B_λ , i.e., the energy per unit wavelength transmitted per unit area, per unit solid angle, per unit time, whose unit is $\text{J}\cdot\text{s}^{-1}\cdot\text{m}^{-2}\cdot\text{sr}^{-1}\cdot\text{m}^{-1}$. For the same temperature T , the peak frequency (wavelength) of the $B_\nu - \nu$ (or $B_\nu - \lambda$) curve is not equal to that of the $B_\lambda - \lambda$ (or $B_\lambda - \nu$) curve. For $T = 2.73$ K, the peak of the $B_\nu - \nu$ and $B_\lambda - \lambda$ curves are approximately 1.6 and 1 mm, respectively.

Fig. 10.11 The cosmic background radiation curve measured by COBE (based on the first results in 1990), the corresponding blackbody temperature is $T \cong 2.735$ K



dipole moment. [Intuitively, as the Earth is going across the “ocean” of background radiation, the radiation in front of it should be stronger than that behind it, which gives rise to the dipole anisotropy.] Therefore, the anisotropy of about one part per thousand obtained by COBE (and the previous ground observations) is not only reasonable, but also it can be used conversely to determine the precise velocity of the Earth relative to the isotropic reference frame (the cosmic rest frame), and the result is about $369 \text{ km}\cdot\text{s}^{-1}$.¹⁴ After this anisotropy is subtracted out, one finds that the anisotropy (mainly the quadrupole anisotropy) when photons are decoupled is only about 10^{-5} . This tiny anisotropy is very critical for understanding the formation of the large scale structure of the universe (e.g., galaxies), see 7 below; when it was first discovered by COBE in 1992, this suddenly became the headline news all over the world. The leaders of the COBE project, G. F. Smoot and J. C. Mather were awarded the 2006 Nobel Prize in Physics for the discoveries of the blackbody spectrum and the anisotropy of the CMB.

Besides its intensity represented by the temperature of the blackbody spectrum, the CMB radiation as electromagnetic radiation also exhibits **polarization**. The CMB polarization can be decomposed into two components, dubbed an **E-mode** and a **B-mode** [for the details of the decomposition, the reader may refer to, e.g., Chap. 10 of Dodelson and Schmidt (2020)].¹⁵ E-modes can be produced by the interaction between photons and free electrons (such as Compton scattering). However, if the radiation is isotropic, the polarization will be equal in all directions, and the overall effect is still unpolarized. In fact, the tiny anisotropy of the CMB temperature we just mentioned plays a crucial role here, which allows the scattering process to

¹⁴ This is the average speed of the Earth relative to the isotropic reference frame, which is also the speed of the Sun relative to the isotropic reference. Since the Earth orbits the Sun with a speed of about 30 km/s, the actual speed of the Earth at each time of the year can be obtained by considering the correction due to this relative motion between the Earth and the Sun.

¹⁵ For the CMB polarization one only considers the polarization patterns on the celestial sphere.

produce polarization. Thus, the spectrum of the E-mode polarization is smaller than the anisotropy spectrum of the CMB. On the other hand, the spectrum of the B-mode polarization is even smaller than that of the E-mode polarization. There are two types of B-mode polarization, the first one is caused by the **gravitational lensing** of E-modes (gravitational lensing is the effect that, due to the deflection of light by gravitational fields, massive bodies such as galaxies behave similar to convex glass lenses); the second one is caused by the **primordial gravitational waves** produced in the early universe. The E-mode polarization and the first type of B-mode polarization have been detected, while the B-modes produced by primordial gravitational waves have not been found yet. The latter, once detected, will provide a powerful tool we can use to see through the early universe, and open a new window for gravitational-wave astronomy (see Sect. 7.9.4).

Since it takes some amount of time for the light emitted by a galaxy to reach the Earth (see Fig. 10.8), from the observations for bright galaxies and quasars one can obtain information of the universe earlier than the present time t_0 . The CMB data carries information way earlier than that (no galaxy was formed yet when photons decoupled), which is highly valuable for the study of cosmology. The observation of the CMB is regarded as the most powerful support to the standard model. One of the drawbacks of a once strong competitor of the standard model—the steady state model—is that it cannot provide a persuasive explanation for the background radiation, and hence it has stepped down from the stage of history since 1965.

7. Structure formation.

The basic premises of the standard model are the large scale spatial homogeneity and isotropy. On a smaller scale, the universe presents a hierarchical structure: there exists stars, galaxies, galaxy clusters and superclusters. A generally accepted idea is that the complicated structure today originates from the extremely weak **density fluctuation** (also called **perturbation**) $\delta\rho/\rho$ in the very early universe, where ρ is the average density, and $\delta\rho$ is the difference between the density at a point and ρ . Gravity has the effect of amplifying the density fluctuation: if $\delta\rho/\rho > 0$ (density is higher than the average density) somewhere, then the matter there will contract under the action of gravity, which leads to a higher density fluctuation. J. H. Jeans has established the corresponding theory for static fluids in 1902, and E. M. Lifshitz proposed the theory for the density fluctuation being amplified in an expanding universe in 1946. Based on these theories, all kinds of models regarding structure formation have been put forward. The early models (in the 1970s) considered that baryons are the largest contributors to the matter in the universe, which leads to serious troubles. Later, after the concept of non-baryonic dark matter was posed (see Sect. 10.3.2 and Chap. 15), two theories of structure formation, namely the **hot dark matter model** and **cold dark matter model**, appeared accordingly [see Longair (2008) for details]. In the hot dark matter model, the formation of structures has a top-down scenario: the superclusters are formed first, and then they break into galaxy clusters and galaxies hierarchically. In contrast, the formation of structures in the cold dark matter model has a bottom-up scenario: the galaxies are formed first, and then galaxy clusters and superclusters are formed hierarchically. As to the

Table 10.1 Chronicle of the evolution of the universe

t	T	$k_B T$	Main events
0.01 s	10^{11} K	10 MeV	A large amount of ν (and $\bar{\nu}$), γ , (e, e^+) and a small amount of p, n are in thermal equilibrium
1 s	10^{10} K	1 MeV	ν (and $\bar{\nu}$) decouple; A large amount of γ , (e, e^+) and a small amount of p, n are in thermal equilibrium
14 s	3×10^9 K	0.3 MeV	(e, e^+) annihilate rapidly
>100 s	< 10^9 K	<0.1 MeV	1. Primordial nucleosynthesis. The products are ^4He (about 25%), H (about 75%) and a tiny amount of ^2H , ^3He , ^7Li 2. (e, e^+) are all annihilated, there remains a small amount of electrons for balancing the charge of protons
10^5 years	3000 K	0.3 eV	Neutral atoms are synthesized; photons decouple and become the background radiation
10^9 years			Structure formation

origin of the primordial perturbation, previously people could only treat it as a pre-assigned initial condition. Nowadays, as the inflationary model has been generally accepted (the basic idea is that the very early universe once experienced a dramatically accelerating exponential expansion in a very short period of time, see Chap. 15 in Volume II), the primordial perturbation can be completely explained by inflation. The cold dark matter model now has achieved great success and is now the favored model. More precisely, the cold dark matter model with the inflationary model offering the “seeds” of the primordial perturbation has became the most widely accepted theory of structure formation. Some even consider it as the fourth cornerstone of the modern cosmology (the first three are the consensual cosmic expansion, primordial nucleosynthesis and the CMB). However, there are still people who have different opinions.

In the end of this subsection, to help readers remember, we roughly summarize a few important periods in the history of the universe’s evolution in Table 10.1.

It should be pointed out that, in the above description of the universe’s evolution, our understanding for the time after $t = 1$ s is relatively reliable. However, for $t < 1$ s, we do not have a description for the early universe with such a high credibility, since any “fossil” from that time has undetermined factors.

10.3.2 The Dark Matter Problem

There only exist three possibilities for the RW metric, $k = 1$, $k = 0$ and $k = -1$. As we have seen in Sect. 10.1.3, the first one is a closed universe, while the latter two are open universes. Which one does our universe really belong to? Is it closed or open?

The answer of course closely relies on astronomical observations. In this subsection we will have some theoretical discussion and introduce some observational results.

It follows from $H \equiv \dot{a}/a$ and (10.2.16) that $H^2 = 8\pi\rho/3 - k/a^2$. Adding back the physical constants G and c (see Appendix A for the details of adding constants), we have

$$H^2 = \frac{8\pi G\rho}{3} - \frac{kc^2}{a^2}. \quad (10.3.8)$$

Define the **critical density**

$$\rho_C := \frac{3H^2}{8\pi G}, \quad (10.3.9)$$

then

$$\rho = \rho_C + \frac{3kc^2}{8\pi Ga^2}. \quad (10.3.10)$$

Thus, $k = 0$ corresponds to $\rho = \rho_C$, and $k = \pm 1$ correspond to $\rho \leq \rho_C$. That is to say, if the mass density ρ of the universe is greater than the critical density ρ_C , then it is a closed universe ($k = 1$), otherwise it is an open universe. This conclusion can be understood intuitively as follows: all kinds of particles are scattered in high speeds at the big bang, and will gradually slow down due to the gravitational effect. If the gravity is strong enough, their speed will gradually decrease to zero and start to accelerate in the opposite direction, which means they will eventually gather back together again (corresponding to a universe that first expands and then contracts); if the gravity is not that strong, although the particles will still keep slowing down, they will never come to a stop and turn around (corresponding to a universe that expends forever). This may be likened to a rocket launched from the Earth: it will eventually fall down with an acceleration if the initial speed is less than some critical value (escape velocity), while if the initial speed is large enough it will leave forever and never come back. The strength of gravity depends on the mass density ρ of the universe, and thus one can expect that there exists a critical value ρ_C , and the universe is closed if and only if $\rho > \rho_C$. Define the density parameter

$$\Omega := \frac{\rho}{\rho_C}, \quad (10.3.11)$$

then Ω can be interpreted as the density with ρ_C as the unit, and hence we can say that the universe is closed if and only if $\Omega > 1$. Notice that ρ_C itself is also a function of t . It follows from (10.3.11) and (10.3.9) that Ω can be expressed as

$$\Omega = \frac{8\pi G\rho}{3H^2}. \quad (10.3.12)$$

Once the present values of the Hubble parameter H_0 and the mass density ρ_0 are measured, then it can be determined from the above equation whether our universe is closed or not. Assume for now that the main contents of the universe are presented

mainly in the form of galaxies. Suppose the present number density of the galaxies is n , and the average mass of the galaxies is \bar{M} , then $\rho_0 = n\bar{M}$. Suppose the luminosity density of the universe per unit volume is \mathcal{L} (called the **luminosity density**), and the average luminosity of the galaxies is \bar{L} , then $\mathcal{L} = n\bar{L}$. Plugging this into $\rho_0 = n\bar{M}$ yields

$$\rho_0 = \mathcal{L} \frac{\bar{M}}{\bar{L}}, \quad (10.3.13)$$

where \bar{M}/\bar{L} is called the **average mass-to-light ratio** of the galaxies. Let ρ_{C0} and Ω_0 represent the present values of ρ_C and Ω , respectively, then

$$\Omega_0 = \frac{\rho_0}{\rho_{C0}} = \frac{8\pi G}{3H_0^2} \mathcal{L} \frac{\bar{M}}{\bar{L}}, \quad (10.3.14)$$

where (10.3.9) and (10.3.13) are used in the second equality. There is already a relatively reliable observational value for \mathcal{L} . Plugging the observational values of \mathcal{L} and H_0 into the above equation, we can get the relation between Ω_0 and the average mass-to-light ratio \bar{M}/\bar{L} . The actual measurement is performed on a galaxy, and the result is only the mass-to-light ratio M/L for this galaxy; only if the galaxy is highly representative can we plug M/L into (10.3.14) as \bar{M}/\bar{L} and get a relatively good result. The masses of different galaxies vary enormously (they may differ by several orders of magnitudes), while the differences in their mass-to-light ratios are much smaller. This is one of the merits of substituting (10.3.12) with (10.3.14) (for $t = t_0$). Now we introduce the dynamical method (which considers the gravitational effect of the mass) of measuring the mass of a spiral galaxy (e.g., the Milky Way). Besides its random motion, a star in a spiral galaxy also undergoes revolution (orbital motion) around the galactic center with the gravity of the galaxy as the centripetal force. To simplify the discussion, we assume that the galaxy has spherical symmetry. From Newton's theory of gravity we know that

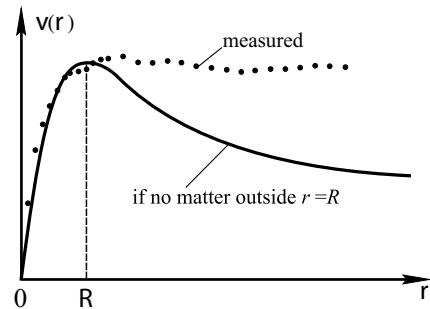
$$v^2(r) = \frac{GM(r)}{r}, \quad (10.3.15)$$

where $v(r)$ is the speed of the orbital motion of a star at a distance r from the center, and $M(r)$ is the mass of the galaxy within the radius r . The curve $v(r)$ is called the **rotation curve** of the galaxy. The rotation curve for many galaxies has been measured. Let R be the r at which the galaxy's luminosity disappears, then $M(R)$ represents the mass of the luminous matter in the galaxy. Plugging the mass-to-light ratio measured in this way into (10.3.14) yields the contribution of the luminous matter to Ω_0 :

$$\Omega_0 (\text{luminous matter}) < 1\% \quad (\text{tends to } 0.5\%). \quad (10.3.16)$$

This indicates that the contribution from all the luminous matter to the mass density is less than one percent of the critical mass. Moreover, $M(R)$ is also way less than

Fig. 10.12 The rotation curve of a galaxy (sketch)



the total mass of the galaxy. If there is no mass outside $r = R$, then it follows from (10.3.15) that the curve for $v(r)$ should decrease as $r^{-1/2}$ starting from $r = R$. However, the rotation curves of plenty of galaxies has the following common property: they first increase steeply from the galactic center, and then extend almost horizontally until very far away from R where it is incapable of measurement,¹⁶ as shown in Fig. 10.12. This indicates that there is a spherical “dark halo” outside the luminous part of a spiral galaxy, formed by non-luminous **dark matter**, whose radius is a lot greater than R , and the mass of this dark halo is 3–10 times as much as the mass of the luminous part. There are also other types of galaxies other than spiral galaxies, e.g., elliptical galaxies. Evidence has indicated that there also exists a considerable amount of dark matter in elliptical galaxies.

Considering that there exists a large space between galaxies, it is very likely that a large amount of matter is there. People have also applied a similar dynamical method for measuring galaxy clusters. [Assume that the Viral theorem holds, then there is a formula similar to (10.3.15)]. The result gives

$$\Omega_0 \text{ (galaxy cluster)} \cong 10\% \sim 30\%. \quad (10.3.17)$$

This confirms that, apart from the galaxies, there is a large amount of dark matter in a galaxy cluster. Since the above results are based on some hypotheses that are not completely conformed yet, and since there are only about 5% of the galaxies in the universe that belong to big clusters of galaxies, we cannot claim that the Ω_0 of the universe can be represented by (10.3.17) (although circumstantial evidence has been found). However, one can at least conclude that the mass of dark matter in the universe is way more than that of luminous matter.

If we take (10.3.17) as the contribution of all the matter in the universe to Ω_0 , we would draw the conclusion that the universe is far from being closed. However, the inflationary model proposed in 1981 (see Chap. 15 in Volume II) suggests that Ω_0 may be very close (or even equal to) 1; this is supported by some measurements and

¹⁶ Each point of the curve is measured from the frequency shift of rays emitted by stars or neutral gas clouds. These stars and gas clouds serve as test particles. It is hard to find a test particle when r is much greater than R .

analyses. As the inflationary model is now widely accepted, how can we coordinate the result $\Omega_0 \cong 1$ and (10.3.17)? Before 1998, in order to avoid the contradiction with $\Omega_0 \cong 1$, people had to think that the distribution of the galaxies and galaxy clusters is far from the total matter distribution of the universe: besides the matter associated with galaxies and galaxy clusters, there might also be about 80% of the matter that is not clustering, or even smoothly distributed in the universe. Note that so far we have only considered the Einstein equation without the Λ -term. The important progress on the measurement of the cosmological constant Λ in 1998 made people believe that one should use the Einstein equation with the Λ -term when discussing cosmology problems. The key point is, besides the contribution Ω_{M0} coming from matter (including luminous matter and dark matter), Ω_0 also has a contribution Ω_Λ from the cosmological constant. The contributions from Ω_Λ and Ω_{M0} roughly has a seventy-thirty ratio, and together they give $\Omega_0 \cong 1$. For details, see Sect. 10.3.3.¹⁷

10.3.3 The Cosmological Constant Problem and the Λ CDM Model

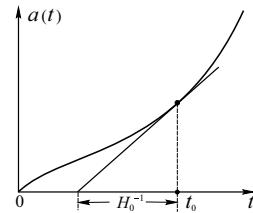
Ever since Einstein introduced the cosmological constant in 1917, the status of Λ has experienced several ups and downs. Although Einstein himself abandoned Λ after 1923, it was still valued by many people until the 1950s. One of the reasons is that the early measurement of H_0 by Hubble was excessively large, and the existence of a positive Λ could avoid the age of the universe being too small. Here is a qualitative explanation. From (10.2.35) we can see that the existence of the cosmological constant is equivalent to adding an “energy-momentum tensor” $-\lambda g_{ab}/8\pi$ to the universe. Comparing this with the energy-momentum tensor of a perfect fluid

$$T_{ab} = (\rho + p)U^a U^b + pg_{ab},$$

we can see that the Λ -term can be considered as a “perfect fluid” with the equation of state $-\rho = p = -\Lambda/8\pi$. If this is the only contributor, then (10.2.18) becomes $3\ddot{a} = a\Lambda$, and for $\Lambda > 0$ we have $\ddot{a} > 0$. Thus, contrary to the matter field, a positive cosmological constant Λ provides a repulsive force unlike the usual attractive gravitational force and makes the universe undergo an accelerating expansion. In the early universe, the energy density ρ of the radiation and matter is very large, and its gravity is stronger than the repulsive effect, which leads to a decelerating expansion. Then, ρ decreases as the universe expands, and the expansion will have a constant rate when ρ is small enough and the gravity counterbalances the repulsive force (note

¹⁷ The important discovery in 1998 is that the present universe is experiencing an accelerating expansion. People once regarded the Λ -term as the cause of this accelerating expansion. However, difficulties still exists. The mechanism of the accelerating expansion of the universe has became one of the biggest puzzles in cosmology or even in fundamental physics, called the “dark energy” problem, see Chap. 15 in Volume II.

Fig. 10.13 The curve of $a(t)$ in the model with $\Lambda > 0$. The age of the universe $t_0 > H_0^{-1}$



that Λ is fixed). As ρ keeps decreasing, the universe will turn into an accelerating expansion when the repulsive force is stronger than gravity. By choosing a suitable model, one can obtain the result that the present universe is undergoing an accelerating expansion, and so the measured H_0^{-1} is less than t_0 instead of greater than t_0 (see Fig. 10.13). So the “age paradox” of the universe can be resolved or at least relieved.

However, the situation turned around in the 1950s. On the one hand, newer measurements indicated that value of H_0 is about 1/8 of that measured by Hubble. On the other hand, the development of the modern theory of stellar evolution also made the age of stars a lot smaller than the estimated values in the 1930s. As a consequence, the “age paradox” disappeared, and Λ became unnecessary again. Nevertheless, after having three ups and downs, nowadays the necessity of Λ has revived once again. The cosmological constant has influenced not only cosmology but also many other areas of physics, and its significance has been generally recognized. However, people are still facing difficulties regarding the cosmological constant. One aspect of it is related to the vacuum energy in quantum field theory, which leads to the **cosmological constant problem for physicists**. Another aspect is from the perspective of astronomers. Now we will introduce the cosmological the constant problem for astronomers; the cosmological constant problem for physicists will be introduced in Chap. 15.

A major concern of astronomers is whether or not a nonzero Λ can be obtained from the observations. The existence of Λ affects the evolution of the universe. After the Λ -term is added into Einstein’s equation, (10.2.16) and (10.2.17) should be modified as follows:

$$\frac{3(\dot{a}^2 + k)}{a^2} = 8\pi\rho + \Lambda, \quad (10.3.18)$$

$$\frac{2\ddot{a}}{a} + \frac{\dot{a}^2 + k}{a^2} = -8\pi p + \Lambda. \quad (10.3.19)$$

Applying (10.3.18) to the present time t_0 yields

$$H_0^2 = \frac{8\pi\rho_0}{3} + \frac{\Lambda}{3} - \frac{k}{a_0^2}. \quad (10.3.20)$$

Following the definition of Ω_0 , one can define

$$\Omega_{\Lambda 0} := \frac{\Lambda}{3H_0^2}, \quad (10.3.21)$$

then (10.3.20) can be written as

$$1 = \Omega_{M0} + \Omega_{\Lambda 0} - \frac{k}{a_0^2 H_0^2}, \quad (10.3.22)$$

Ω_{M0} stands for the contribution from matter to Ω_0 (note that the present contribution of radiation is negligible). Through observation, the following questions should be answered: Do we need a nonzero $\Omega_{\Lambda 0}$ to assure that the above equation holds? If so, what is the value of $\Omega_{\Lambda 0}$? This may be referred to as the **cosmological constant problem for astronomers**.

The speed \dot{a} and the acceleration \ddot{a} of the universe's evolution depends on the overall effect of Ω_M representing the gravity and Ω_Λ representing the repulsive force. Astronomers usually use a dimensionless quantity to represent the deceleration $-\ddot{a}$; what they care about is the present value of the (dimensionless) **deceleration parameter** defined as follows:

$$q_0 := - \left(\frac{a}{\dot{a}^2} \ddot{a} \right)_{t_0}. \quad (10.3.23)$$

Considering that in the present day we have $\rho + 3p \cong \rho$, we can plug the present values of (10.3.18) and (10.3.19) (take $p = 0$) into (10.3.23) and obtain

$$q_0 = \frac{1}{2} \Omega_{M0} - \Omega_{\Lambda 0}. \quad (10.3.24)$$

The above equation intuitively reflects the fact that “ Ω_{M0} leads to a deceleration and a positive $\Omega_{\Lambda 0}$ leads to an acceleration”. The direct measurement of q_0 has been going on for decades. One of the main difficulties is how to choose a suitable object to measure (a distance indicator). The clustered galaxies used to be taken as the measured objects, but they have a shortcoming—their own evolution will bring undetermined factors into the measurements. What we need is a distance indicator that is not sensitive to evolution. Later, people found that Type Ia (also denoted by 1a) supernovae can serve as ideal distance indicators, and the measurements of them have became an active subject. A large number of observational results on high redshift Type Ia supernovae published since 1998 [e.g., Riess et al. (1998); Perlmutter et al. (1999)] has attracted huge attention internationally. These results indicate with a high confidence level that: ① the cosmological constant is nonzero, and is positive; ② unlike what people used to think, the present universe is experiencing an accelerating expansion (the effect of $\Omega_{\Lambda 0}$ exceeds the effect of $\Omega_{M0}/2$, which leads to $q_0 < 0$). Furthermore, the combination of these results and the observational results of the anisotropy of the CMB provides the following quantitative result [Aghanim et al. (2018)]:

$$\Omega_{M0} = 0.311 \pm 0.006, \quad \Omega_{\Lambda0} = 0.689 \pm 0.006, \quad (10.3.25)$$

This indicates that: ① $\Omega_{M0} + \Omega_{\Lambda0} \cong 1$, and thus [according to (10.3.22)] $k \cong 0$, i.e., the present universe is very close to flat, which agrees with the prediction of the inflationary model introduced later in Chap. 15; ② $\Omega_{\Lambda0}$ is not only far from innocuous, but also dominates the contribution to Ω_0 . The ratio of $\Omega_{\Lambda0}$ and Ω_{M0} is about seventy-thirty. The accelerating expansion of the universe is regarded as one of the most groundbreaking discoveries in the 20th century. S. Perlmutter, B. P. Schmidt and A. G. Riess are awarded the 2011 Nobel Prize in Physics for this discovery. Further interpretation of this accelerating expansion leads to the topic of “dark energy”, which will be introduced in Chap. 15.

The above result also have support from another aspect: although the cold dark matter (CDM) model is the most successful model in the theory of structure formation, the CDM model based on $\Omega_{M0} = 1$ cannot fit the observations. In contrast, if a positive Λ is included in the theory with $\Omega_{M0} \cong 0.3$ and $\Omega_{\Lambda0} \cong 0.7$, the resulting model, dubbed the **Λ CDM model** fits the observation result very well.

As a relatively simple model developed based on the FLRW model, the Λ CDM model is also often called the **standard cosmological model** or the **concordance cosmological model** in modern literature due to its success in explaining the observation results. However, this is not the end of the story as the standard model is still facing challenges and can be further extended. We will come back to this in Chap. 15.

Exercises

- ~10.1. Verify that the curvature tensor ${}^{(3)}R_{abc}{}^d$ of the metric in (10.1.12) satisfies ${}^{(3)}R_{ab}{}^{cd} = 2\bar{R}^{-2}\delta_a{}^{[c}\delta_b{}^{d]}$.
- 10.2. Show that the world line of an isotropic observer is a geodesic. Hint: from the expressions for the Christoffel symbols below (10.2.5) and (5.7.2), this is almost obvious.
- 10.3. Derive the formula (10.2.8) for cosmological redshift from the following steps:
 - Show that any null geodesic $\eta(\beta)$ (where β is an affine parameter) has $d\omega/d\beta = -K^a K^b \nabla_a Z_b$, where

$$K^a \equiv (\partial/\partial\beta)^a, \quad Z^a \equiv (\partial/\partial t)^a, \quad \omega \equiv -g_{ab}Z^a K^b.$$

- (b) Show that $\nabla_a Z_b = (\dot{a}/a)h_{ab}$, where h_{ab} is the metric on the surface of homogeneity induced by g_{ab} , and $\dot{a} \equiv da/dt$.

Hint: first show that $\nabla_a Z_b$ is a spatial tensor field, i.e., $Z^a \nabla_a Z_b = 0 = Z^b \nabla_a Z_b$, and then show that the results of applying both sides of $\nabla_a Z_b = (\dot{a}/a)h_{ab}$ on $(\partial/\partial x^i)^a (\partial/\partial x^j)^b$ ($i, j = 1, 2, 3$) are the same.

- (c) Using the results of (a) and (b), derive that $d\omega/\omega = -da/a$, which gives (10.2.8).
- 10.4. The present age of the universe is the time it takes for the evolution from $a = 0$ to $a_0 \equiv a(t_0)$. Given any value of a , one can talk about the time it takes for the scale factor of the universe to evolve to this value, which is called the age of the universe corresponding to this value of a . Therefore, the age t can be regarded as a function of a .
- (a) Starting from (10.2.30) and (10.2.26), show that the age function of a matter-dominated universe with $\Lambda = 0$ is given by the following three equations:

$$\text{for } \Omega_0 = 1, \quad t = \frac{2}{3H_0^{-1}} \left(\frac{a}{a_0} \right)^{3/2},$$

for $\Omega_0 > 1$,

$$t = H_0^{-1} \left\{ \frac{\Omega_0}{2(\Omega_0 - 1)^{3/2}} \cos^{-1} \left[1 - 2(1 - \Omega_0^{-1}) \frac{a}{a_0} \right] - \frac{1}{\Omega_0 - 1} \left[\Omega_0 \frac{a}{a_0} - (\Omega_0 - 1) \left(\frac{a}{a_0} \right)^2 \right]^{1/2} \right\},$$

for $\Omega_0 < 1$,

$$t = H_0^{-1} \left\{ \frac{-\Omega_0}{2(1 - \Omega_0)^{3/2}} \cosh^{-1} \left[1 + 2(\Omega_0^{-1} - 1) \frac{a}{a_0} \right] + \frac{1}{1 - \Omega_0} \left[\Omega_0 \frac{a}{a_0} + (1 - \Omega_0) \left(\frac{a}{a_0} \right)^2 \right]^{1/2} \right\}.$$

- (b) Derive the expressions for the present age t_0 of the universe in the cases $\Omega_0 = 1, \Omega_0 > 1, \Omega_0 < 1$ from the above three equations.
- ~10.5. Show that the Einstein equation with the Λ -term does not admit a solution having a flat metric even if there is no matter field ($T_{ab} = 0$). Hint: find the relation of R and T from the Einstein equation with the Λ -term, so that the R in the equation can be eliminated. Then, it is easy to see that R_{ab} cannot vanish when $T_{ab} = 0$.

References

- Aghanim, N. et al. (2020), ‘Planck 2018 results. VI. Cosmological parameters’, *Astron. Astrophys.* **641**, A6. [arXiv:1807.06209](https://arxiv.org/abs/1807.06209).
- Cyburt, R. H., Fields, B. D., Olive, K. A. and Yeh, T.-H. (2016), ‘Big bang nucleosynthesis: 2015’, *Rev. Mod. Phys.* **88**, 015004. [arXiv:1505.01076](https://arxiv.org/abs/1505.01076).
- Dodelson, S. and Schmidt, F. (2020), *Modern Cosmology*, Academic Press, London.
- Ellis, G. F. R. (1989), The expanding universe: a history of cosmology from 1917 to 1960, in D. Howard and J. Stachel, eds, ‘Einstein and the History of General Relativity’, Birkhäuser, Boston, pp. 367–431.
- Follin, B., Knox, L., Millea, M. and Pan, Z. (2015), ‘First detection of the acoustic oscillation phase shift expected from the cosmic neutrino background’, *Phys. Rev. Lett.* **115**, 091301. [arXiv:1503.07863](https://arxiv.org/abs/1503.07863).
- Hicks, N. J. (1965), *Notes on Differential Geometry*, Van Nostrand, Princeton.
- Kolb, E. W. and Turner, M. S. (1990), *The Early Universe*, Addison-Wesley Publishing Company, Redwood City.

- Longair, M. S. (2008), *Galaxy Formation*, Springer-Verlag, Berlin.
- Peebles, P. J. E. (1993), *Principles of Physical Cosmology*, Princeton Press, Princeton.
- Peebles, P. J. E. and Ratra, B. (2003), ‘The cosmological constant and dark energy’, *Rev. Mod. Phys.* **75**, 559–606. [arXiv:astro-ph/0207347](https://arxiv.org/abs/astro-ph/0207347).
- Perlmutter, S. et al. (1999), ‘Measurements of Ω and Λ from 42 high redshift supernovae’, *Astrophys. J.* **517**, 565–586. [arXiv:astro-ph/9812133](https://arxiv.org/abs/astro-ph/9812133).
- Riess, A. G. et al. (1998), ‘Observational evidence from supernovae for an accelerating universe and a cosmological constant’, *Astron. J.* **116**, 1009–1038. [arXiv:astro-ph/9805201](https://arxiv.org/abs/astro-ph/9805201).
- Riess, A. G. et al. (2022), ‘A comprehensive measurement of the local value of the Hubble constant with $1 \text{ km s}^{-1} \text{ Mpc}^{-1}$ uncertainty from the Hubble space telescope and the SH0ES team’, *Astrophys. J. Lett.* **934**(1), L7. [arXiv:2112.04510](https://arxiv.org/abs/2112.04510).
- Rindler, W. (1982), *Introduction to Special Relativity*, Clarendon Press, Oxford.
- Steigman, G., Schramm, D. N. and Gunn, J. E. (1977), ‘Cosmological limits to the number of massive leptons’, *Phys. Lett. B* **66**, 202–204.
- Steigman, G. (2012), ‘Neutrinos and big bang nucleosynthesis’, *Adv. High Energy Phys.* **2012**, 268321. [arXiv:1208.0032](https://arxiv.org/abs/1208.0032).
- Di Valentino, E., Mena, O., Pan, S., Visinelli, L., Yang, W., Melchiorri, A., Mota, D. F., Riess, A. G. and Silk, J. (2021), ‘In the realm of the Hubble tension—a review of solutions’, *Class. Quant. Grav.* **38**(15), 153001. [arXiv:2103.01183](https://arxiv.org/abs/2103.01183).
- Wald, R. M. (1984), *General Relativity*, The University of Chicago Press, Chicago.
- Weinberg, S. (2008), *Cosmology*, Oxford University Press, Oxford.
- Zyla, P. A. et al. (2020), ‘Review of particle physics’, *PTEP* **2020**(8), 083C01.

Appendix A

The Conversion Between Geometrized and Nongeometrized Unit Systems

When discussing systems of units, one should pay attention to the distinction between a quantity and a number. Besides quantity equations, what is more commonly used are numerical-valued equations. The form of a numerical-value equation depends on the system of units, and thus when memorizing a physical formula we should also remember in which system of units it holds. Since the speed of light in vacuum and the gravitational constant are frequently involved in relativity, setting their numerical values to 1 (i.e., $c = G = 1$) will simplify the equations a lot, and the corresponding system of units is called the **geometrized unit system**. However, geometrized units are inconvenient for calculating the numerical values of physical quantities. Now we will introduce the conversion of physical equations between systems of geometrized units and non-geometrized units (e.g., SI).

To avoid confusion, we will use bold and regular letters to represent quantities and numbers, respectively (only in this appendix). A non-geometrized unit system usually takes the time T , length L and mass M as the base quantities in mechanics. In the geometrized unit system, since $c = G = 1$, only one of these three quantities can be chosen arbitrarily, and hence we can say that there is only one base quantity. For instance, one can choose time as the base quantity and choose s (second) as its unit (base unit). However, the essence of $c = 1$ is to take the speed of light as a unit of speed, and so one can consider the speed V as a base quantity in the geometrized unit system, and the speed of light is a base unit. Similarly, $G = 1$ implies that the gravitational constant G is also a base quantity in the geometrized unit system. Therefore, one can also say that there are three base quantities, i.e., T , V and G . In fact, the number of base quantities in the same system of units is flexible, and one can choose them according to the specific context. Suppose A is an arbitrary quantity whose numerical value in the International System of Units (SI) and the geometrized unit system are A and A' , respectively, then their ratio

$$\chi \equiv \frac{A'}{A} \tag{A.1}$$

is called the **conversion factor** of A between the two systems. The reason that χ is not equal to 1 is that the units of T , L and M are different in the two systems. The units of T , L and M in SI are the **s**, **kg** and **m**, respectively. In the geometrized unit system, the only certain thing is that $c = G = 1$, while the units of T , L and M are somewhat flexible. For the convenience of comparing the two systems, we stipulate the time unit in the geometrized system to also be **s**. Under this stipulation, one can determine the geometrized units of L and M using $c = G = 1$, and there is no longer any flexibility (see Optional Reading A.1). According to dimensional analysis, the relation between a derived unit and the base units is given by the dimensional equation:

$$[A] = [T]^{\tau} [L]^{\lambda} [M]^{\mu}. \quad (\text{A.2})$$

When we only care about the conversion between the geometrized unit system and SI (or the Gaussian unit system), the $[T]^{\tau}$ in this system can be ignored since the time unit is the same in the two systems:

$$[A] = [L]^{\lambda} [M]^{\mu}. \quad (\text{A.3})$$

What the dimensional equation describes is how a derived unit changes with a change of the base units. For instance, once we treat the $[L]$ and $[M]$ in the above equation as multiples of the units of the base quantities L and M , respectively, then $[A]$ represents the corresponding multiple of the unit of the derived quantity A . In this interpretation, all of $[A]$, $[L]$ and $[M]$ represent numbers, and (A.3) should be interpreted as a numerical-value equation. Changes of the units of L and M lead to corresponding changes of the units of the velocity V and the gravitational constant G , their relations obey

$$[V] = [L], \quad [G] = [L]^3 [M]^{-1}. \quad (\text{A.4})$$

Combining the equation above with (A.3) yields

$$[A] = [V]^{\lambda+3\mu} [G]^{-\mu}. \quad (\text{A.5})$$

Suppose the multiple of the units of L and M when we turn from SI to the geometrized system are $[L]$ and $[M]$, respectively, then the multiple of V and G are $[V]$ and $[G]$ in (A.4); the multiple of A is $[A]$ in (A.5), and comparing with (A.1) we can see that $\chi = [A]$. The speed of light and the true value of the gravitational constant in SI are c and G , which are both 1 in the geometrized system, and hence $[V] = 1/c$, $[G] = 1/G$. Plugging these into (A.5) yields $[A] = c^{-\lambda-3\mu} G^{\mu}$. Therefore,

$$\chi = c^{-\lambda-3\mu} G^{\mu}. \quad (\text{A.6})$$

Equations (A.1) and (A.6) indicate that to find the numerical value of A in SI from its value A' in the geometrized system, we only have to know the dimensional exponents λ and μ of A with respect to the base quantities L and M , which can be easily derived or looked up.

Example 1 Find the expression of the Schwarzschild radius in SI from its expression $r'_S = 2M'$ in the geometrized system.

Solution Suppose A is a quantity whose numerical value in SI is $A \equiv r_S/M$, then its numerical value in the geometrized system is $A' \equiv r'_S/M' = 2$. It follows from $[A] = [L][M]^{-1}$ that $\lambda = 1$, $\mu = -1$, and it follows from (A.6) that $\chi = c^2 G^{-1}$. Then, from $\chi \equiv A'/A$ we have $A' = c^2 G^{-1} A$, and hence $r_S/M \equiv A = c^{-2} G A' = 2c^{-2} G$. Therefore, the expression for the Schwarzschild radius in SI is $r_S = 2GM/c^2$. ■

Example 2 Convert the form of the timelike normalization condition $Z'^a Z'_a = -1$ in the geometrized system to its form in SI.

Solution $Z'^a Z'_a = -1$ is equivalent to $g'_{\mu\nu}(dx'^{\mu}/d\tau')(dx'^{\nu}/d\tau') = -1$. Choose x^μ and x^ν as length coordinates, then it follows from $ds^2 = g_{\mu\nu} dx^\mu dx^\nu$ that $[g_{\mu\nu}] = 1$, $[dx^\mu/d\tau] = [T]^{-1}[L]$, and hence for a quantity $dx^\mu/d\tau$ we have $\lambda = 1$, $\mu = 0$, $\chi = c^{-1}$ and

$$g'_{\mu\nu}(dx'^{\mu}/d\tau')(dx'^{\nu}/d\tau') = c^{-2} g_{\mu\nu}(dx^\mu/d\tau)(dx^\nu/d\tau).$$

Therefore, $g_{\mu\nu}(dx^\mu/d\tau)(dx^\nu/d\tau) = -c^2$, i.e., $Z^a Z_a = -c^2$. ■

Example 3 In Sect. 10.2.2 we used the expression for the angular frequency of a photon in the geometrized system $\omega' = dt'/d\beta'$ (β is the affine parameter of the photon's world line). Find its form in SI.

Solution First we should figure out the dimension of β . The wave 4-vector of the photon is $K' = (\partial/\partial\beta')^a$. It follows from $K'^a = \omega'(\partial/\partial t')^a + k'^a$ that $[K^a] = [k^a]$.¹ Since $k^a = k^i(\partial/\partial x^i)^a$, where k^i are the components of the wave 3-vector, and $[k^i] = [L]^{-1}$, we have $[K^i] = [L]^{-1}$, and thus $[\beta] = [L]^2$. The expression $\omega' = dt'/d\beta'$ can be written as $\omega' d\beta'/dt' = 1$. Let $A' \equiv \omega' d\beta'/dt'$, then $[A] = [T]^{-2}[L]^2$. Hence $\lambda = 2$, $\mu = 0$, $\chi = c^{-2}$, $A' = c^{-2} A$, i.e., $\omega' d\beta'/dt' = c^{-2} \omega d\beta/dt$, and so $\omega = c^2 dt/d\beta$. ■

In general relativity we often encounter tensors like g_{ab} , $R_{abc}{}^d$, R_{ab} and R . When it comes to the conversion of units, we will need to know the dimensions of these quantities. For the convenience of the conversion, first we prove the following conclusions (note that the indices can be unbalanced in an dimension equation):

$$\begin{aligned} (1) \quad [g_{ab}] &= [L]^2, & (2) \quad [\nabla_a \omega_b] &= [\omega_b], & (3) \quad [R_{abc}{}^d] &= 1, \\ (4) \quad [R_{abcd}] &= [L]^2, & (5) \quad [R_{ac}] &= 1, & (6) \quad [R] &= [L]^{-2}. \end{aligned} \quad (\text{A.7})$$

¹ The dimension of a vector can be defined as the dimension of the real number (quantity) obtained by acting the vector on a dimensionless scalar field. Similarly one can define the dimension of a dual vector and a tensor.

Proof (1) Since the essence of $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$ is $g_{ab} = g_{\mu\nu}(dx^\mu)_a(dx^\nu)_b$, we have $[g_{ab}] = [ds^2] = [L]^2$. (The readers who feel confused about this may consider it from the perspective of the components. When x^μ and x^ν are both length coordinates, from $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$ and $[ds^2] = [L]^2$ we can see that $[g_{\mu\nu}] = 1$. Then it follows from $g_{ab} = g_{\mu\nu}(dx^\mu)_a(dx^\nu)_b$ that $[g_{ab}] = [L]^2$. One should notice that, unlike $[g_{ab}]$ which is absolute, $[g_{\mu\nu}]$ relies on the dimension of the coordinates involved.)

$$(2) [\nabla_a \omega_b] = [\partial_a \omega_b] = [(dx^\mu)_a(dx^\nu)_b \partial \omega_\nu / \partial x^\nu] = [(dx^\nu)_b][\omega_\nu] = [\omega_b].$$

(3) $\nabla_a \nabla_b \omega_c - \nabla_b \nabla_a \omega_c = R_{abc}{}^d \omega_d$. Considering that $[\nabla_a \nabla_b \omega_c] = [\omega_c]$, we have $[R_{abc}{}^d \omega_d] = [\omega_d]$, and thus $[R_{abc}{}^d] = 1$.

$$(4) [R_{abcd}] = [g_{de} R_{abc}{}^e] = [L]^2.$$

$$(5) [R_{ac}] = [g^{bd} R_{abcd}] = [L]^{-2} [L]^2 = 1.$$

$$(6) [R] = [g^{ac} R_{ac}] = [L]^{-2}.$$

□

Example 4 Find the form of Einstein's equation in SI from its form $R'_{ab} - R' g'_{ab}/2 = 8\pi T'_{ab}$ in the geometrized system.

Solution For simplicity (and without loss of generality), take a perfect fluid as an example. The energy momentum tensor of a perfect fluid is

$$T'_{ab} = (\rho' + p') U'_a U'_b + p' g'_{ab}.$$

Since the dimensions of the terms being summed are the same, we only have to consider how does the equation $R'_{ab} = 8\pi p' g'_{ab}$ transform. $[R_{ab}] = 1$ leads to $[R'_{ab}] = [R_{ab}]$. Also $[pg_{ab}] = [M][L]^{-1}[T]^{-2} \cdot [L]^2 = [M][L][T]^{-2}$, and hence for the quantity $p\mathbf{g}_{ab}$ we have $\lambda = \mu = 1$, $\chi = c^{-4}G$. Thus, $p' g'_{ab} = c^{-4} G p g_{ab}$, and so $R_{ab} = 8\pi c^{-4} G p g_{ab}$. Therefore, the form of Einstein's equation in SI is

$$R_{ab} - \frac{1}{2} R g_{ab} = \frac{8\pi G}{c^4} T_{ab}. \quad (\text{A.8})$$

■

Until now we only talked about the conversion of the units in mechanics. Although we used SI as an example for non-geometrized systems, the discussion can also be applied to the Gaussian system. However, when electromagnetism is involved, one needs to add a fourth base quantity, then the difference between SI and the Gaussian system will be revealed. The fourth base quantity in SI is the electric current I , whose base unit is the ampere; the fourth base quantity in the Gaussian system is the permittivity ϵ , whose base unit is the permittivity of the vacuum ϵ_0 (and thus the number $\epsilon_0 = 1$). Correspondingly, the equations in the geometric system that has electromagnetic quantities also have two forms, which may be called the “geometrized SI” and “geometrized Gaussian system”. Besides the basic requirement $c = G = 1$, the geometrized Gaussian system also requires $\epsilon_0 = 1$, while the geometrized SI stipulates that the unit of electric current is the ampere. To match the international literature, we adopt the geometrized Gaussian system for all the equations in this text that have electromagnetic quantities. The equations that do not have electromagnetic

quantities have the same form in the two geometrized systems. It is not difficult to see that the method above can be applied to both the conversion from the geometrized Gaussian system to the Gaussian system and that from the geometrized SI to SI. For instance, it is straightforward for the reader to convert the form of the RN line element in the geometrized Gaussian system

$$ds'^2 = -\left(1 - \frac{2M'}{r'} + \frac{Q'^2}{r'^2}\right)dt'^2 + \left(1 - \frac{2M'}{r'} + \frac{Q'^2}{r'^2}\right)^{-1}dr'^2 + r'^2(d\theta'^2 + \sin^2\theta'd\varphi'^2), \quad (\text{A.9})$$

to the following form in the Gaussian system:

$$ds'^2 = -\left(1 - \frac{2GM}{c^2r} + \frac{GQ^2}{c^4r^2}\right)c^2dt^2 + \left(1 - \frac{2GM}{c^2r} + \frac{GQ^2}{c^4r^2}\right)^{-1}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2). \quad (\text{A.10})$$

To facilitate lookup, we list some of the equations involving electromagnetic quantities in the form of geometrized SI as follows (the equation numbers without * are the corresponding equations in the geometrized Gaussian system):

$$\partial^a F_{ab} = -\epsilon_0^{-1} J_b, \quad (6.6.10^*)$$

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0}, \quad \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}, \quad \vec{\nabla} \cdot \vec{B} = 0, \quad \vec{\nabla} \times \vec{B} = \mu_0 \vec{j} + \frac{\partial \vec{E}}{\partial t}. \quad (6.6.12^*)$$

$$T_{ab} = \epsilon_0(F_{ac}F_b{}^c - \frac{1}{4}\eta_{ab}F_{cd}F^{cd}), \quad (6.6.28^*)$$

$$T_{ab} = \frac{\epsilon_0}{2}(F_{ac}F_b{}^c + {}^*F_{ac}{}^*F_b{}^c), \quad (6.6.28'^*)$$

$$T_{00} = \frac{\epsilon_0}{2}(E^2 + B^2) \quad w_i = -T_{i0} = \epsilon_0(\vec{E} \times \vec{B})_i, \quad i = 1, 2, 3, \quad (\text{no number})$$

$$ds'^2 = -\left(1 - \frac{2M}{r} + \frac{Q^2}{4\pi\epsilon_0 r^2}\right)dt'^2 + \left(1 - \frac{2M}{r} + \frac{Q^2}{4\pi\epsilon_0 r^2}\right)^{-1}dr'^2 + r'^2(d\theta'^2 + \sin^2\theta d\varphi'^2), \quad (8.4.26^*)$$

$$F_{ab} = -\frac{Q^2}{4\pi\epsilon_0 r^2}(dt)_a \wedge (dr)_a, \quad \text{or} \quad A_a = -\frac{Q}{4\pi\epsilon_0 r}(dt)_a. \quad (8.4.27^*)$$

All of the $1/2\pi$ in (8.8.7) are changed to $2\epsilon_0$, and all of the factors 2 in (8.8.8) and (8.8.9) are changed to $8\pi\epsilon_0$. The $-2\pi J_\mu$ in Exercise 8.10 is changed to $-\frac{1}{2}\epsilon_0 J_\mu$.

[Optional Reading A.1]

This optional reading further introduces the geometrized system (still restricted to mechanics). Question: How large are the units of the length \mathbf{L} and mass \mathbf{M} (as quantities)? It will be convenient for answering this question if we choose \mathbf{T} , \mathbf{V} and \mathbf{G} as the base quantities. The dimension equations of \mathbf{L} and \mathbf{M} with respect to these three base quantities are

$$[L] = [T][V], \quad [M] = [T][V]^3[G]^{-1}. \quad (\text{A.11})$$

Let L_G and L_I represent the number obtained by measuring the same length using the length units in the geometrized system and SI respectively, then $L_G/L_I = [L]$. Note that the time units in the geometrized system and non-geometrized systems are respectively 1 and $c = 3 \times 10^8$, which means $[L] = 1/c$, and hence the above equation gives $L_I = cL_G$. Thus,

$$\text{length unit in the geometrized system} = c \times \text{length unit in the SI} = 3 \times 10^8 \text{ m}. \quad (\text{A.12})$$

Similarly, it follows from the second equation in (A.11) and $[G] = 1/G$ (where the number $G = 6.67 \times 10^{-11}$) that

$$\begin{aligned} \text{mass unit in the geometrized system} &= \frac{c^3}{G} \times \text{mass unit in the SI} \\ &= \frac{(3 \times 10^8)^3}{6.67 \times 10^{-11}} \times \text{kg} = 4 \times 10^{35} \text{ kg}. \end{aligned} \quad (\text{A.13})$$

On the other hand, when we do not need to change the units of V and G , it is quite beneficial to regard the geometrized system as having only one base quantity T . In this case, we can view three originally different quantities—time, length and mass—as the same type of quantity. The “key” to identifying them is to regard 1 s, 3×10^8 m and 4×10^{35} kg as equal, i.e.,

$$1 \text{ s} = 3 \times 10^8 \text{ m} = 4 \times 10^{35} \text{ kg}, \quad (\text{A.14})$$

and so a quantity has either no unit or a unit of s [or a power of s (which is a quantity)]. For instance, ① the Earth has a speed $v \cong 10^{-3}$ relative to the center of the Milky Way, this numerical value ($\ll 1$) strongly indicates that the Earth’s speed is so slow that the observations of the universe made by an Earth’s observer can be regarded as those made by an (imaginary) observer at the center of the Milky Way. ② In the geometrized system, the distance from the Earth to the Sun is 480 s, which indicates intuitively that it takes eight minutes for light to travel from the Sun to the Earth. ③ The Earth’s radius and mass in the geometrized system are $R_\oplus \cong 2 \times 10^{-2}$ s and $M_\oplus \cong 1.5 \times 10^{-11}$ s; $M_\oplus \ll R_\oplus$ indicates that the gravitational field on the Earth’s surface is so weak that Newton’s theory is a good approximation in most of the cases.

[The End of Optional Reading A.1]

The geometrized unit system is very convenient for general relativity. For a quantum theory that does not involve gravity, a **natural unit system** is frequently used, in which $c = \hbar = 1$. Depending on the field involved, sometimes one can also set a third physical constant to 1. For example, k_B (the Boltzmann constant) is set to 1 when thermodynamics is involved, m_e (the value of the electron mass) is set to 1 when atomic physics is involved, m_p or m_n (the value of the proton or neutron mass) is set to 1 when nuclear physics is involved, and $G = 1$ when gravity is involved (e.g., a theory of quantum gravity). The unit system with $G = c = \hbar = 1$ is also called the **Planck unit system**. Now we discuss the conversion between the Planck system and SI. Compared with the geometrized system, the Planck system has an additional constraint $\hbar = 1$ besides $G = c = 1$, which prevents one from choosing the unit of time (and thus all the quantities) arbitrarily. Therefore, we should start from (A.2) [instead of (A.3)], and change (A.4) to

$$[V] = [T]^{-1}[L], \quad [G] = [T]^{-2}[M]^{-1}[L]^3. \quad (\text{A.15})$$

Combining (A.2) and (A.15) yields

$$[A] = [V]^{\lambda+3\mu} [G]^{-\mu} [T]^{\lambda+\mu+\tau}. \quad (\text{A.16})$$

It is not difficult to show that the “unique” quantity with time dimension constructed by the speed of light, gravitational constant, and the reduced Planck constant is the Planck time t_P , whose numerical value in SI is $t_P = (G\hbar/c^5)^{1/2} \sim 10^{-43}$ (s), where c , G and \hbar are the numerical values of the speed of light, the gravitational constant, and the reduced Planck constant in SI, respectively. The values of these three quantities in SI are all 1, and hence $[V] = 1/c$, $[G] = 1/G$, $[T] = 1/t_P$. Suppose $\tilde{\chi}$ is the conversion factor for the quantity A between SI and the Planck system, then it follows from (A.16) that

$$\tilde{\chi} = c^{-\lambda-3\mu} G^\mu t_P^{-(\lambda+\mu+\tau)} = c^{-\lambda-3\mu} G^\mu (G\hbar/c^5)^{-(\lambda+\mu+\tau)/2}. \quad (\text{A.17})$$

Example 5 The relation of the energy E and the frequency ν in the Planck form reads $E' = 2\pi\nu'$. Find its form in SI.

Solution Suppose $A \equiv E/\nu$, then $[A] = [E][\nu]^{-1} = [T]^{-1}[M][L]^2$, and hence $\tau = -1$, $\mu = 1$, $\lambda = 2$. Plugging this into (A.17) yields

$$\tilde{\chi} = c^{-5} G (G\hbar/c^5)^{-1} = \hbar^{-1}.$$

Thus, $A' = \tilde{\chi} A = \hbar^{-1} A$, i.e., $E'/\nu' = \hbar^{-1} E/\nu$. Hence, it follows from $E'/\nu' = 2\pi$ that $E/\nu = 2\pi\hbar$, or $E = 2\pi\hbar\nu$. ■

Example 6 The “unique” quantity with mass dimension constructed by the speed of light, the gravitational constant, and the reduced Planck constant is the Planck mass m_P , whose numerical value in the Planck system is $m'_P = 1$. Find its numerical value m_P in SI.

Solution From $[m_P] = [M]$ we can see that $\tau = \lambda = 0$, $\mu = 1$. Plugging this into (A.17) yields $\tilde{\chi} = c^{-3} G (G\hbar/c^5)^{-1/2} = (\hbar c/G)^{-1/2}$, and hence $m'_P = \tilde{\chi} m_P = (\hbar c/G)^{-1/2} m_P$. Thus, it follows from $m'_P = 1$ that $m_P = (\hbar c/G)^{1/2}$. Plugging in the numerical values $\hbar = 10^{-34}$, $c = 3 \times 10^8$, $G = 6.67 \times 10^{-11}$ we obtain $m_P = 2.1 \times 10^{-8}$ kg. ■

Exercises

- A.1. The form of the relation of the energy, mass and momentum of a point mass in the geometrized system reads $E'^2 = m'^2 + p'^2$. Find its form in SI.
- A.2. Find the form of hydrostatic equation in SI from its form $dp'/dr' = -\rho'm'/r'^2$ in the geometrized system.

- A.3. The form of the Euler equation of non-relativistic hydrodynamics in the geometrized system reads $-\nabla p' = \rho'[\partial \vec{u}'/\partial t' + (\vec{u}' \cdot \nabla) \vec{u}']$. Find its form in SI.
- A.4. Show that the equations $U'^a = \gamma(Z'^a + u'^a)$ [see (6.3.30)] and $\omega' = -K'^a Z'_a$ [see (6.6.42)] have the same form in the geometrized system and SI.
- A.5. The geometrized system in some literature [e.g., Sachs and Wu (1977)] is defined by $c = 1 = 8\pi G$ (the time unit is still s). Find the units of length and mass in this system (in this problem, one should regard the gravitational constant as one of the base quantities of the geometrized system).
- A.6. The “unique” quantity with length dimension constructed by the speed of light, the gravitational constant, and the reduced Planck constant is the Planck length l_P , whose numerical value in the Planck system is $l'_P = 1$. Find its numerical value l_P in SI.

Reference

Sachs, R. K. and Wu, H. (1977), *General Relativity for Mathematicians*, Springer-Verlag, New York.

Conventions and Notation

Note on Conventions

- (1) Starting from Sect. 2.6, this work has adopted the abstract index notation to represent tensors. For instance, v^a represents a vector, where the Latin letter a , called an abstract index, plays a similar role to the \rightarrow in the commonly used notation \vec{v} . Do not interpret v^a as the a th component of v^a . When talking about the components we use Greek letters as the indices (called component indices or concrete indices); for example, v^μ represents the μ th component of the vector v^a . There is only one exception: a vector v^a in a 4-dimensional spacetime has three spatial components, for which we will use the most commonly used convention, i.e., using v^i (where $i = 1, 2, 3$) to represent the i th component of v^a . Although this violates the stipulation of “using Latin letters to represent the abstract indices”, it is convenient in many ways. In order to distinguish from the abstract indices a, b, c, d, e, \dots , we only use Latin letters starting from i (usually i, j, k) as the labels for the spatial components. Practice has shown that this can effectively avoid confusion. For more details about the index notation, see Sect. 2.3.
- (2) This work adopts the signature convention $- + + +$ for the metric of 4-dimensional spacetime.
- (3) The definitions of the Riemann tensor $R_{abc}{}^d$ and the Ricci tensor R_{ab} have various conventions in the literature. This work follows the conventions of Wald (1984).

Notation List

$\{ \}$	Set. First appears in Sect. 1.1. E.g., $X = \{1, 4, 5.6\}$ stands for the set formed by the real numbers 1, 4 and 5.6.
\mathbb{R}	The set of real numbers. First appears in Sect. 1.1.
\mathbb{N}	The set of natural numbers. First appears in Sect. 1.3.
S^n	n -dimensional sphere.
$\forall x$	For all x . First appears in Sect. 1.1.
\exists	There exists. First appears in Sect. 1.1.
\in	Belongs to. First appears in Sect. 1.1. E.g., $x \in X$ stands for “ x belongs to the set X ”, i.e., x is an element of X .
\notin	Does not belong to. First appears in Sect. 1.1.
\subset	Contained in. First appears in Sect. 1.1. E.g., $A \subset X$ stands for “ A is contained in the set X ”, i.e., A is a subset of X .
\subsetneq	Contained in but not equal to. First appears in Sect. 1.1. E.g., $A \subsetneq X$ stands for “ A is contained in but not equal to the set X ”, i.e., A is a proper subset of X .
\cup	Union (see Definition 2 of Sect. 1.1).
\cap	Intersection (see Definition 2 of Sect. 1.1).
$-$	Difference of sets, e.g., $A - B$ stands for the difference of the sets A and B (see Definition 2 of Sect. 1.1).
$-A$	Complement of A (see Definition 2 of Sect. 1.1).
\emptyset	Empty set. First appears in Sect. 1.1.
$:=$	Defined as. First appears in Sect. 1.1.
\equiv	Identical to or denoted by. First appears in Sect. 1.1. E.g., $A \equiv B \cup C$ means “denote $B \cup C$ by A ”.
\approx	Approximately equal to.
\Rightarrow	Implies (if ... then), e.g., $A \Rightarrow B$ stands for “if A then B ”.
\Leftrightarrow	Equivalent to (if and only if).
\times	Cartesian product (see Definition 3 of Sect. 1.1).
\square	Q.E.D. (Denotes the end of a proof, aligned to the right.)
\mathbb{R}^n	The set of n -tuples (x^1, \dots, x^n) of real numbers, i.e., $\mathbb{R}^n = \mathbb{R} \times \dots \times \mathbb{R}$ (n factors in total).
\otimes	Tensor product (see Definition 2 of Sect. 2.4).
\rightarrow	Map. First appears in Sect. 1.1. E.g., $f : X \rightarrow Y$ stands for “the map from X to Y ”.
$f[A]$	Suppose $f : X \rightarrow Y$, $A \subset X$, then the image of A under the action of f is denoted by $f[A]$ in order to distinguish it from the image $f(x)$ of $x \in X$ under f .
\mapsto	Maps to (image of a function), e.g., suppose $f : X \rightarrow Y$, $x \in X$, $y \in Y$, then $x \mapsto y$ stands for “the image of x is y ”.
\circ	Composite map. First appears in Sect. 1.1. E.g., $\phi \circ \psi$ stands for the composite map of ϕ and ψ (ψ after ϕ).
(X, \mathcal{T})	The topological space with X as the base set and \mathcal{T} as the topology (see Definition 2 of Sect. 1.2 and the following paragraph).
\mathcal{T}_u	Usual topology (see Example 3 of Sect. 1.2).
C^r	The first r derivatives exist and are continuous.
C^∞	Smooth (derivatives of all orders exist and are continuous).

\bar{A}	The closure of the set A (see Definition 8 in Sect. 1.2).
$i(A)$	The interior of the set A (see Definition 9 in Sect. 1.2).
∂A or ∂A	The boundary of the set A (see Definition 10 in Sect. 1.2).
T_2 space	Hausdorff space (see Definition 3 in Sect. 1.3).
$\dim V$	The dimension of V .
V^*	The dual space of the vector space V . First appears in Sect. 2.3.
V_p	The tangent space at a point p in the manifold. First appears in Sect. 2.2.
V_p^*	The dual space of the vector space V_p .
\vec{E}	3-dimensional (spatial) vector. (Use an arrow above the letter instead of boldface, which is used elsewhere, see the ω later.)
e_μ or $(e_\mu)^a$	The μ th basis vector in the chosen basis $\{(e_\mu)^a\}$.
$e^{\mu*}$ or $(e^\mu)_a$	The μ th dual basis vector of the basis $\{(e_\mu)^a\}$.
$\frac{\partial}{\partial x^\mu}$ or $(\frac{\partial}{\partial x^\mu})^a$	The μ th coordinate basis vector field. First appears in Example 2 of Sect. 2.2.
dx^μ or $(dx^\mu)_a$	The μ th dual coordinate basis vector field. First appears after (2.3.8).
$\mathcal{T}_V(k, l)$	The set of all the tensors of type (k, l) on a vector field V . First appears after Example 1 of Sect. 2.4.
\mathcal{F}_M or \mathcal{F}	The set of all the smooth functions on a manifold M (see Definition 5 of Sect. 2.1).
$\mathcal{F}_M(k, l)$	The set of all the smooth tensor fields of type (k, l) on a manifold M . First appears in Definition 1 of Sect. 3.1.
\tilde{A}	The transpose of a matrix A .
$[u, v]$	The commutator of the vector fields u and v (see Definition 10 of Sect. 2.2).
C	Contraction. E.g., suppose $T \in \mathcal{T}_V(2, 2)$, then C_2^1 stands for the contraction between the first upper index and the second lower index. First appears in Remark 2 of Sect. 2.4. In terms of abstract indices it is expressed by $C_2^1 T \equiv T^{ab}_{ca}$.
δ or δ_{ab}	Euclidean metric (see Definition 8 of Sect. 2.5).
η or η_{ab}	Minkowski metric (see Definition 8 of Sect. 2.5).
δ^a_b	Identity map. First appears in the paragraph around (2.6.4).
$g(u, v)$	The result of acting the metric tensor g on u and v . First appears in Definition 1 of Sect. 2.5. Same as $g_{ab}u^a v^b$.
$T_{(abc)}$	The total symmetrization over the indices a, b, c [see (2.6.13) for definition].
$T_{[abc]}$	The total antisymmetrization over the indices a, b, c [see (2.6.14) for definition].
∇_a	Derivative operator (see Definition 1 of Sect. 3.1).
∂_a	The ordinary derivative operator in a coordinate system [see (3.1.9) for definition]. In special relativity it refers to the ordinary derivative operator in an inertial coordinate system, satisfying $\partial_a \eta_{bc} = 0$.
Γ^a_{bc}	The Christoffel symbol of a derivative operator in a coordinate system (see Definition 2 of Sect. 3.1).
$\Gamma^\mu_{\nu\sigma}$	The components of the Christoffel symbol Γ^a_{bc} in a coordinate system.
\exp	Exponential map (see Optional Reading 3.3.1).
ϕ^*	The pullback map induced by the map ϕ (see Definitions 1 and 3 of Sect. 4.1).

ϕ_*	The pushforward map induced by the map ϕ (see Definitions 2 and 4 of Sect. 4.1).
$\mathcal{L}_v T^{\cdots \dots}$	The Lie derivative of the tensor field $T^{\cdots \dots}$ along a vector field v^a (see Definition 1 of Sect. 4.2).
ω	Differential form (field) (in boldface, indices are omitted). For instance, ϵ is the abbreviation for the volume element $\epsilon_{a_1 \dots a_n}$ (n -form field).
${}^*\omega$	The dual differential form of ω (see Definition 1 of Sect. 5.6).
$\Lambda(l)$	The set of all the l -forms on a vector space V . First appears after Theorem 5.1.2.
$\Lambda_M(l)$	The set of all the l -form fields on a manifold M . First appears in Definition 3 of Sect. 5.1.
$\Lambda_p(l)$	The set of all the l -forms at a point p (i.e., on V_p). First appears in the beginning of Sect. 5.6.
d	Exterior differentiation operator (see Definition 3 of Sect. 5.1), e.g., $d\omega$ stands for the exterior differentiation of the differential form ω .
\wedge	Wedge product (see Definition 2 of Sect. 5.1).
ω_μ^v	Connection 1-form, also denoted by $\omega_\mu^v{}_a$. First appears in (5.7.4).
R_μ^ν	Curvature 2-form, also denoted by $R_{ab\mu}^\nu$. First appears in (5.7.7).
\mathcal{R}	Reference frame. First appears in Sect. 6.1.1.
$\frac{D}{d\tau}$	The covariant derivative along a curve $G(\tau)$, e.g., $\frac{Dv^a}{d\tau}$ means the same as $T^b \nabla_b v^a$ [T^b stands for the tangent vector of $G(\tau)$].
$\frac{D_F}{d\tau}$	The Fermi derivative along a curve $G(\tau)$ (see Definition 1 of Sect. 7.3).
Re	Take the real part.
Im	Take the imaginary part.
$(\epsilon_\mu)^a$	The μ th basis vector in a null tetrad. First appears in Sect. 8.7.

Reference

Wald, R. M. (1984), *General Relativity*, The University of Chicago Press, Chicago.

Index

A

Absolute simultaneity, 176
surface of, 176, 178, *See also* Exercise 6.21

Absolute time, 176, 178

Abstract index notation, 55

Abundance, 500, 508

Accumulation point, 16

Active viewpoint, 107

Active viewpoint (language), 401

Adapted coordinate system, 111

Affine parameter, 84

Age of the universe, 499, *See also* Exercise 10.4

Angular frequency, 228, 249, 305, 413

Angular momentum, 411

Angular velocity 2-form, 254

Antisymmetric tensor, 60, 130

Arc length, 49, 344
of a geodesic, 85, 347
parameter, 50

Arcwise connected, 136

Associated volume element, 142

Asymmetry of matter and antimatter, 507

Atlas, 20

Average mass-to-light ratio, 518

Axisymmetry, 355

B

Baryon, 508, 510, 515

Baryonic dark matter, 515

Basis
dual, 38, 43
dual coordinate, 40
orthonormal, 47

Bianchi identity, 94, 283, 366

Big bang, 495, 504

Bijection, 4

Binary star, 318

Binding energy, 509

Birkhoff's theorem, 347, 355, 358, 453, 456

Black hole, 317, 434, 439, 452

Blackbody radiation (theorem, curve, spectrum), 213, 512

Blueshift
Doppler, 234

B-mode, 514

Boost, 53, 115, 169

Boundary, 12, 139, 146

C

Calibration curve, 174

Cartan's first equation of structure, 154

Cartan's second equation of structure, 154, 361

Cartesian coordinate system, 52

Cartesian product, 2

Charge density, 219

Chart, 20

Christoffel symbol, 72, 79, 97, 153, 268
contracted, 99
equivalent definition of, 153
of a static spherically symmetric line element, 342
of the Schwarzschild metric, 345
of the Vaidya metric, 379, *See also* Exercise 3.4

Clock synchronization, 170

Closed differential form, 133, 226

Closed set, 12

Closed universe, 479

Closure, 12

Cluster of galaxies, 468, 515
 CMB, *see* cosmic microwave background radiation
 CMB polarization, 321, 514
 COBE, 513
 Commutation relation, 366
 Commutator, 33, 75, 112, 276
 Comoving observer, *see* observer
 Comoving reference frame, 212, 474
 Compactness, 13
 Compatible, 20, 141, 142, 147
 Complement, 2
 Complete vector field, 37, 110, 113
 Component index notation, 56
 Composite map, *see* map
 Concrete index notation, *see* component index notation
 Congruence, 276
 Conjugate points, 87, 281, 347
 Connected manifold, 136
 Connected topological space, 12, 136
 Connection, 77, 81
 Connection 1-form, 153, 361
 Connection coefficients, 153
 Conservation equation, 209
 Conservation of baryon number, 508
 Conservation of mass, 193
 Conserved quantity, 195
 Constant map, *see* map
 Continuity equation, 209, 214
 Continuous map, *see* map
 Contraction, 44
 Contravariant index, 60
 Contravariant vector, 60
 Conversion factor, 528
 Convex neighborhood, 484
 Coordinate
 basis, 27
 basis vector, 27
 components, 27
 line, 29, 111, 168, 253, 260, 267, 331, 341, 355, 398
 patch, 20
 singularity, 439, 442
 system, 20
 time, 171
 transformation, 20, *See also* Exercise 9.9
 Coordinate clock, 281
 Coordinate condition, 398
 Coordinates, 20
 Coriolis force, 262
 Cosmic microwave background radiation, 493, 511, 513

Cosmic rest frame, 474, 514
 Cosmic time, 481
 Cosmological constant, 501
 Cosmological scale, 468
 Cosmology, 467
 Covariant derivative, 74, 79
 Covariant index, 60
 Covariant vector, 60
 Covector, 120
 Critical density, 436, 517
 Curvature singularity, 442
 Curvature tensor, 92, 100, 154
 Curve, 28
 Cyclic identity, 94
 Cylindrical symmetry, 357, 370

D

Dark matter, 515, 519
 Deceleration parameter, 522
 Degeneracy pressure, 432
 Degenerate electron gas, 432
 Degenerate “metric”, 124, 179
 De Morgan’s law, 2
 Density fluctuation, 515
 Derivative operator, 67
 associated with a metric, 79
 covariant, 72
 non-commutativity of, 93, 100, 247
 ordinary, 72
 torsion-free, 69
 Diffeomorphism, 22
 group, 337, *See also* one-parameter group of diffeomorphisms
 Difference, 2
 Differentiable manifold, 19
 Differential form, 132
 Differential structure, 21
 Dipole anisotropy, 513
 Discrete topology, 7
 Distance, 3, 339, 344, 387
 Doppler effect, 233, 462
 Dual basis, *see* basis
 Dual coordinate basis, *see* basis
 Dual differential form, 149, 217, 249, 254
 Duality rotation, 404
 Dual space, 37
 Dual vector, 37
 Dust, 214, 218, 286, 462

E

Eddington-Finkelstein coordinates, 457, *See also* Exercise 9.13

- Einstein (field) equation, 282
 linearized, 288
 vacuum, 285, 341, 396, 453
 with source, 285, 349, 399, 422
- Einstein's elevator, 267, 275, 279
- Einstein's equation with source, *see* Einstein (field) equation
- Einstein equivalence principle (EEP), 267
- Einstein-Maxwell equations, 350, 370
- Einstein spacetime, 405
- Einstein's static universe, 503
- Einstein tensor, 284
 linearized, 288
- Electric field, 217, 350, 354
- Electromagnetic 4-potential, 226, 247, 288
 equation of motion of, 226, 248
 gauge freedom of, 226, 289
 of the RN solution, 351
- Electromagnetic field
 nonnull, 350, 354, 358, 372
 null, 350, 359, 369, 372
- Electromagnetic field tensor, 220, 223, 225, 245
 in the NP formalism, 369, 372
- Electromagnetic wave, 226, 248, 348
- Electron degeneracy pressure, 432
- Electrovacuum, 349, 355, 371
- Elliptical polarization, 232
- Embedded submanifold, 119
- Embedding, 119
- Embedding diagram, 454
- E-mode, 514
- Empty set, 1
- Energy conservation, 193, 209, 212
- Energy density, 206, 211, 225, 284, 424, 512
- Energy flux density, 207, 225, 390
- Energy-momentum tensor, 207
 of an electromagnetic field, 225
 of a perfect fluid, 211, *See also* Exercise 6.17, 6.18
- Equivalence principle, 267
- Euclidean group, 357
- Euclidean metric (space), 51
- Euler equation, 214
- Event, 163
- Event horizon, 452, 456, 459
- Exact differential form, 133, 226
- Exponential map, *see* map
- Extension, 34, 92, 349, 440, 445
 Kruskal, 449
- Exterior differentiation, 132
- Extrinsic curvature, 100
- F**
- Faster-than-light, 167
- Fermi energy, 432
- Fermi momentum, 435
- Fermi-Walker derivative, 251
- Fermi-Walker transport, 252
- Finite subcover, 13
- Fluid particle, 212
- Foliation, 469
 leaf, 469
 homogeneous, 471
- 4-acceleration, 187, 201, 244, 246, 250
- 4-current density, 219
- 4-force, 203
- 4-momentum, 200, 389
 conservation of, 201
 density, 208
- 4-potential, *see* electromagnetic 4-potential
- 4-velocity, 195
- 4-velocity field, 211
- Frame, 154, 205, 411
- Frame of reference, *see* reference frame
- Freely falling observer, 263, 462, 464
- Freely falling reference frame, 276
- Free point mass, 178, 202, 242, 247, 263, 272
- Friedmann equation, 495
- Friedmann-Lemaître-Robertson-Walker model, 498
- Future (past)-directed, 177
- G**
- Galaxy, 468, 515
- Galilean coordinates, 178
- Garage paradox, 188
- Gauge freedom (gauge transformation)
 of angular velocity, 255, 257, 258
 of electromagnetic 4-potential, 226, 289, 351
 of general relativity, 402
 of the linearized theory of gravity, 289
- Gaussian normal coordinates, 398
- Gauss's theorem, 148, 209
- General covariance, principle of, 245
- Generalized Riemannian space, 50
- Geodesic, 83, 178, 407
- Geodesic deviation equation, 278, 282, 322
- Geometric optics approximation, 229, 248, 413
- Geometrized unit system, 527
- Gravitational collapse, 317, 434, 439, 456
- Gravitational lensing, 515
- Gravitational mass, 241

Gravitational potential, 178, 239, 275, 294, 425
 Gravitational radiation, 296, 348
 Gravitational redshift, 414, 461, 463
 Gravitational wave, 296
 cross-polarized, 307
 plus-polarized, 307
 polarization modes of, 299, 324
 primordial, 515
 Graviton, 308, 326
 Group, 35

H

Harmonic coordinate condition, 398
 Harmonic function, 398
 Hausdorff space, 14
 Hodge dual, *see* dual differential form
 Homeomorphism, 9
 Homogeneity, 468, 471
 spatial homogeneity, 469
 Homogeneous, 471
 Hubble constant, 488
 Hubble-Lemaître law, 488
 Hubble parameter, 489
 Hubble tension, 500
 Hulse-Taylor binary, 318
 Hydrostatic equilibrium, 426
 Hypersurface, 119
 null, 122, 176, 228, 315, 452
 spacelike, 122, 170, 176, 398
 timelike, 122
 Hypersurface orthogonal, 333, 340, 453

I

Identity map, *see* map
 Incomplete geodesic, 440, 449
 Incomplete vector field, *see* complete vector field
 Indiscrete topology, 7
 Inertial coordinate system, 165, 168
 Inertial coordinate time, 171
 Inertial force, 262, 267
 Inertial mass, 241
 Inertial reference frame, *see* reference frame
 Inextensible integral curve, 35
 Infinitesimal coordinate transformation, 290
 Inflation, 317, 467
 Injection, 4
 Instantaneous observer, *see* observer
 Instantaneous rest (inertial) reference frame
 (observer, coordinate system), 199, 263, 387

Integral curve, 34, 228
 Integral of a function, 145
 Integration on manifolds, 134
 Interior, 12
 Interior Schwarzschild solution, 428
 Intersection, 2
 Intrinsic curvature, 92, 100
 Invariant, 195
 Inverse image, 3
 Inversion, 54
 Isometry, 113, 333, 336, *See also* Exercise 4.12
 Isometry group, 337

Isotropic coordinate system, 397
 Isotropic observer, 471
 Isotropic reference frame, *see* reference frame
 Isotropic spacetime, 471
 Isotropy, 212, 468

J

Jacobi identity, *see* Exercise 2.8

K

Killing equation, 114
 Killing vector field, 113
 Kinnersley metric, 383, 387
 Kruskal extension (coordinates), 446

L

Λ CDM model, 523
 Laser interferometer, 319
 Leaf, 469, 470
 Left-handed system (basis), 136
 Length contraction, 179, 189, 219
 Lie derivative, 110
 Lightlike vector, 48
 LIGO, 319
 Line element, 50
 induced, 84
 Linear approximation, 287
 Linearized Einstein (field) equation, 288
 Linearized Einstein tensor, 288
 Linearized Riemann tensor, 288
 Local inertial frame, 248, 269
 Locally unique, 34, 85
 Local Lorentz frame (system), 269
 Local measurement, 196, 410, 424
 Lorentz contraction, *see* length contraction
 Lorentz covariance, 189, 239
 Lorentz 4-force, 223, *See also* Exercise 6.18

Lorentzian coordinate system, 53, 118, 260, 269, 478
Lorentzian metric, *see* metric
Lorentz transformation, 117, 169
Lorenz gauge, 226
 of linearized gravity, 289, 296
Luminosity density, 518

M

Mach's principle, 240, 469
Macroscopic point, 468
Magnetic field, 217, 350, 351
Manifold, 19
Manifold with boundary, 139
Map, 3
 composite, 4
 constant, 4
 continuous, 5, 9
 exponential, 89
 identity, 20
 one-to-one, 4
 onto, 4
 projection, 124, *See also* Exercise 1.9
Mass conservation, 214
Mass defect, 193
Mass dipole moment, 316
Matter, 493
Matter field, 206
Maxwell's equations, 220, 350
 in NP formalism (source-free), 368
 in NP formalism (with source), 405
Metric, 47
 indefinite, 47
 induced, 122, 125, 147
 Lorentzian, 47
 negative definite, 47
 positive definite, 47
Microwave background, 511
Milky Way, 468
Minimal substitution rule, 246
Minkowski metric (space, spacetime), 53, 166
Mode +, 307, 324
Mode ×, 307, 324
Momentum density, 207, 225, *See also* Exercise 6.17
Momentum flux density, 208, 390

N

Natural coordinates, 3, 19
Natural unit system, 532
Neighborhood, 11

Neutrino background, 508
Neutrino decoupling, 508
Neutron degeneracy pressure, 433
Neutron star, 317, 433
Newman-Penrose formalism, 360
Newtonian gravity
 4-dimensional formulation of, 177
Newtonian limit, 292
Newtonian spacetime, 178
Non-degenerate, 47, 58
Non-locality of the gravitational field energy, 425
Non-rotating observer, 249, 263, 267
Normal coordinates, *see* Gaussian or Riemannian normal coordinates
Normal covector, 120, 228, 315
Normal neighborhood, 90
Normal vector, 121, 228, 315
NP equations, 364, 371
Null curve, 49
Null electromagnetic field, *see* electromagnetic field
Null hypersurface, *see* hypersurface
Null tetrad, 360, 367, 381
Null vector, 48, 360, 380
Number of neutrino species, 511
Numerical-value equation, 527

O

Observer, 164
 comoving, 211
 inertial, 167
 instantaneous, 196
 instantaneous rest, 199
 instantaneous rest inertial, 199, 263, 387
 non-rotating, 249, 263, 267
 stationary (static), 334
One-parameter family of geodesics, 276
One-parameter group of diffeomorphisms, 36, 110, 113
One-parameter group of isometries, 113
One-parameter local group of diffeomorphisms, 37
One-to-one map, *see* map
Onto map, *see* map
Open (sub)set, 7
Open ball, 7
Open cover, 13, 19
Open disk, 8
Open universe, 479
Oppenheimer-Volkoff equation, 426
Orbit, 36, 337

Orbit sphere, 337, 408
 Orientable manifold, 135
 Orientation, 135, 139
 induced, 140, 147
 Orthonormal, 47

P

P.p. curvature singularity, 442
 Parallel (vectors), 31
 Parallel transport, 75, 80
 Parameter, 28, 36
 Parametric representation (equation), 29
 Parametrization, 28, 49
 Passive viewpoint (language), 107, 401
 Perfect fluid, 211, 422
 Perihelion precession, 415
 Perturbation, 515
 Phase, 227
 Photon, 164, 167, 230
 background, 512
 decoupling, 512
 energy, 230
 momentum, 230
 rocket, 389
 Photon gas, 213
 Planck energy (mass), 533
 Planck length, 506, 534
 Planck's law of blackbody radiation, 512
 Planck time, 505, 533
 Planck unit system, 532
 Plane symmetry, 357
 Polarization
 cross-polarized, 307
 plus-polarized, 307
 Polarization direction, 310
 Polarization vector (tensor), 227, 303
 Polytrope, 437
 Pp-wave, 305
 Preimage, 3
 Primordial nucleosynthesis, 508
 Primordial perturbation, 516
 Product topology, 8
 Projection map, *see* map
 Proper coordinate system, 259
 Proper distance, 345
 Proper number density, 219
 Proper subset, 1
 Proper time, 164, 170
 Pseudo-Cartesian coordinates, 53
 Pseudo-Euclidean space, 93
 Pseudo-Riemannian space, 51
 Pseudo-rotation, 255

Pullback map, 105
 Pure radiation field, 380, 388, 391
 Pushforward map, 105

Q

Quadrupole anisotropy, 513
 Quantity equation, 527
 Quantum gravity, 308, 505

R

Radiation, 213, 380, 493
 Radiation gauge, 297
 of linearized gravity, 298
 Radius, 340
 Schwarzschild, 439, 529
 Rate, 170
 Ratio of the densities of the baryons and photons, 510
 Recessional velocity, 488
 Red giant, 431
 Redshift, 453, 461, 488
 cosmological, 490
 Doppler, 234, 463
 gravitational, 414, 463
 Redshift parameter, 414
 Reference frame, 165
 geodesic, 276, 321
 inertial, 168, 172
 isotropic, 213, 471, 480
 static, 334
 stationary, 334
 Reflection, 52, 53
 Regular embedding, 119
 Reissner-Nordström (RN) metric (line element), 354
 Relativistic mass, 191
 Relativistic particle, 507
 Reparametrization, 28
 Rest energy, 193
 Rest mass, 191, 206
 Restriction, 138, 139
 Retarded distance (time), 387
 Revolution, 262
 Ricci flat, *see* Exercise 7.7
 Ricci rotation coefficients, 156, 361
 Ricci tensor, 96, 363
 components in a null tetrad, 363
 first-order (linear) approximation of, 288
 Riemann (curvature) tensor, 92, 153
 of the Schwarzschild metric, 345
 Riemannian normal coordinates, 91
 Riemannian space, 51

- Riemann tensor
linearized, 288
- Right-handed system (basis), 136, 140, 142, 143, 146, 148, 151
- Rigid frame, 156, 360
- Rindler spacetime (coordinates), 442
- Robertson-Walker metric, 483
- Rotation, 52, 116, 254, 262
- Rotation curve, 518
- S**
- Scalar curvature, 96, 285, 364
- Scalar field, 23, 26, 351
- Scalar multiplication, 23
- Scale factor, 484
- Schwarzschild radius, 439, 529
- Semi-plane symmetric, 359
- Sequence, 16
- Setting, 170
- Signature, 47
- Simultaneity
absolute, 176, 178
relative, 175, 176
surface of, 175, 334, 470
- Singularity, 439 >big bang, *see* big bang coordinate, 439, 442
spacetime, 440, 441, 450, 504
theorems, 504, *See also* Exercise 9.9
- Singular spacetime, 441
- Slice, 470
- Slicing, 469
- Smooth, 10, 19, 23, 32, 40, 46
- Space of constant curvature, 476
- Spacelike curve, 49
- Spacelike hypersurface, *see* hypersurface
- Spacelike vector, 48
- Spacetime, 163, *See also* Exercise 5.9
- Spacetime diagram, 164, 172, 462
- Spacetime rotation, 254
- Spatial distance, 344, 387, *See also* Exercise 8.8
- Spatial homogeneity, 470, 471
- Special relativity, 469
- Spherically symmetric metric field, 337
- Spherically symmetric spacetime, 337
- Spin coefficients, 363
- Standard clock, 164, 170
- Standard cosmological model, 498, 515
- Standard model, *see* standard cosmological model
- Static reference frame (observer, spacetime), 334
- Stationary spacetime, 331
- Stokes's Theorem, 139
- Strong equivalence principle (SEP), 272
- Structure formation, 515
- Subset, 1
- Supercluster, 468, 515
- Supernova, 317, 434, 522
- Surface of homogeneity, 471
- Surface of infinite redshift, 453
- Surjection, 4
- Symmetric tensor, 47, 60
- Synchronous gauge, 300
- T**
- T_2 space, 14
- Tangent space (tangent plane), 31, 119, 269
- Tangent vector, 31, 119
- Tensor, 42
multifaceted view of, 43, 56, 322
product, 43, 55
transformation law, 46
- Tensor field, 46
- Tetrad, 205, 252, 360, 367, 381, 411
- 3-acceleration, 187, 195, 202, 241, 261, 278
- 3-current density, 219
- 3 + 1 decomposition, 175, 469
- 3-force, 195, 204
- 3-momentum, 195, 200, 389
conservation of, 201
of a photon, 230
- 3-momentum flux density (tensor), 208, 390
- 3-speed, 198
- 3-velocity, 195, 198, 261, 278
- Tidal acceleration, 275, 278, 296, 322
- Tilde force (effect), 273, 346, 460
- Time, 469
- Time dilation, 181, 461
- Timelike curve, 49
- Timelike hypersurface, *see* hypersurface
- Timelike vector, 48
- Time-orthogonal coordinate system, 334
- Topological space, 7
- Topology, 6
induced, 9
- Torsion 2-form, 155
- Torsion tensor, 155, *See also* Exercise 3.1
- Tortoise coordinate, 378, 447
- Totally (anti)symmetric tensor, 61
- Trace, 44, 95, 350, *See also* Exercise 8.4
- Translation, 36, 52, 53, 116, 168
- Translational invariance, 35, 115, 332, 357
- Transverse gauge, 299

Transverse-traceless (TT) gauge, 298, 303,
See also Exercise 7.9
 Triad, 205, 254, 411
 Trivial manifold, 20, *See also* Exercise 2.2
 True singularity, 439
 Twin paradox, 186

U

Union, 2
 Universe, 467
 Usual topology, 7

V

Vacuum Einstein equation, *see* Einstein
 (field) equation
 Vacuum Schwarzschild solution, 341
 Vaidya metric, 378
 Vector, 24
 (components) transformation law, 28
 Vector field, 32

Vector space, 23
 Volume, 143
 Volume element, 141
 compatible with the metric (orientation),
 142, 147
 induced, 147

W

Wave 3-vector, 228, 305
 Wave 4-vector, 228, 248, 303, 413
 Wavefront, 228
 Weak Equivalence Principle (WEP), 267
 Weber bar, 318
 Wedge product, 130
 Weyl tensor, 96, 363, 374
 component in a null tetrad, 363, 374
 White dwarf, 433
 White hole, 452
 World line, 164
 World sheet, 175, 180, 228, *See also* Exer-
 cise 5.10