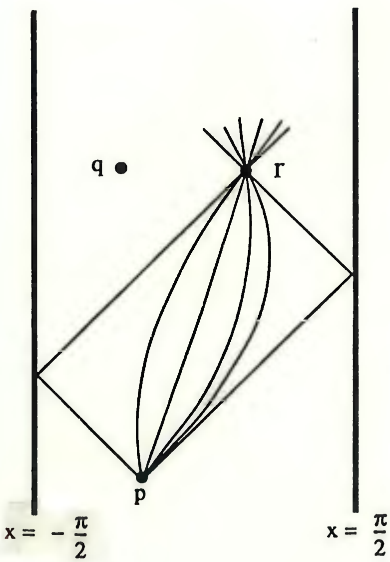
PURE AND APPLIED MATHEMATICS

A Series of Monographs and Textbooks

GLOBAL LORENTZIAN GEOMETRY

Second Edition



John K. Beem
Paul E. Ehrlich
ley

GLOBAL LORENTZIAN GEOMETRY

PURE AND APPLIED MATHEMATICS

A Program of Monographs, Textbooks, and Lecture Notes

MONOGRAPHS AND TEXTBOOKS IN PURE AND APPLIED MATHEMATICS

1. *K. Yano,* Integral Formulas in Riemannian Geometry (1970)
2. *S. Kobayashi,* Hyperbolic Manifolds and Holomorphic Mappings (1970)
3. *V. S. Vladimirov,* Equations of Mathematical Physics (A. Jeffrey, ed.; A. Littlewood, trans.) (1970)
4. *B. N. Pshenichnyi,* Necessary Conditions for an Extremum (L. Neustadt, translation ed.; K. Makowski, trans.) (1971)
5. *L. Narici et al.,* Functional Analysis and Valuation Theory (1971)
6. *S. S. Passman,* Infinite Group Rings (1971)
7. *L. Dornhoff,* Group Representation Theory. Part A: Ordinary Representation Theory. Part B: Modular Representation Theory (1971, 1972)
8. *W. Boothby and G. L. Weiss, eds.,* Symmetric Spaces (1972)
9. *Y. Matsushima,* Differentiable Manifoids (E. T. Kobayashi, trans.) (1972)
10. *L. E. Ward, Jr.,* Topology (1972)
11. *A. Babakhanian,* Cohomological Methods in Group Theory (1972)
12. *R. Gilmer,* Multiplicative Ideal Theory (1972)
13. *J. Yeh,* Stochastic Processes and the Wiener Integral (1973)
14. *J. Barros-Neto,* Introduction to the Theory of Distributions (1973)
15. *R. Larsen,* Functional Analysis (1973)
16. *K. Yano and S. Ishihara,* Tangent and Cotangent Bundles (1973)
17. *C. Procesi,* Rings with Polynomial Identities (1973)
18. *R. Hermann,* Geometry, Physics, and Systems (1973)
19. *N. R. Wallach,* Harmonic Analysis on Homogeneous Spaces (1973)
20. *J. Dieudonné,* Introduction to the Theory of Formal Groups (1973)
21. *I. Vaisman,* Cohomology and Differential Forms (1973)
22. *B.-Y. Chen,* Geometry of Submanifolds (1973)
23. *M. Marcus,* Finite Dimensional Multilinear Algebra (in two parts) (1973, 1975)
24. *R. Larsen,* Banach Algebras (1973)
25. *R. O. Kujala and A. L. Vitter, eds.,* Value Distribution Theory: Part A; Part B: Deficit and Bezout Estimates by Wilhelm Stoll (1973)
26. *K. B. Stolarsky,* Algebraic Numbers and Diophantine Approximation (1974)
27. *A. R. Magid,* The Separable Galois Theory of Commutative Rings (1974)
28. *B. R. McDonald,* Finite Rings with Identity (1974)
29. *J. Satake,* Linear Algebra (S. Koh et al., trans.) (1975)
30. *J. S. Golan,* Localization of Noncommutative Rings (1975)
31. *G. Klambauer,* Mathematical Analysis (1975)
32. *M. K. Agoston,* Algebraic Topology (1976)
33. *K. R. Goodearl,* Ring Theory (1976)
34. *L. E. Mansfield,* Linear Algebra with Geometric Applications (1976)
35. *N. J. Pullman,* Matrix Theory and Its Applications (1976)
36. *B. R. McDonald,* Geometric Algebra Over Local Rings (1976)
37. *C. W. Groetsch,* Generalized Inverses of Linear Operators (1977)
38. *J. E. Kuczkowski and J. L. Gersting,* Abstract Algebra (1977)
39. *C. O. Christenson and W. L. Voxman,* Aspects of Topology (1977)
40. *M. Nagata,* Field Theory (1977)
41. *R. L. Long,* Algebraic Number Theory (1977)
42. *W. F. Pfeffer,* Integrals and Measures (1977)
43. *R. L. Wheeden and A. Zygmund,* Measure and Integral (1977)
44. *J. H. Curtiss,* Introduction to Functions of a Complex Variable (1978)
45. *K. Hrbacek and T. Jech,* Introduction to Set Theory (1978)
46. *W. S. Massey,* Homology and Cohomology Theory (1978)
47. *M. Marcus,* Introduction to Modern Algebra (1978)
48. *E. C. Young,* Vector and Tensor Analysis (1978)
49. *S. B. Nadler, Jr.,* Hyperspaces of Sets (1978)
50. *S. K. Segal,* Topics in Group Kings (1978)
51. *A. C. M. van Rooij,* Non-Archimedean Functional Analysis (1978)
52. *L. Corwin and R. Szczarba,* Calculus in Vector Spaces (1979)
53. *C. Sadosky,* Interpolation of Operators and Singular Integrals (1979)

54. *J. Cronin,* Differential Equations (1980)
55. *C. W. Groetsch,* Elements of Applicable Functional Analysis (1980)
56. *I. Vaisman,* Foundations of Three-Dimensional Euclidean Geometry (1980)
57. *H. I. Freedan,* Deterministic Mathematical Models in Population Ecology (1980)
58. *S. B. Chae,* Lebesgue Integration (1980)
59. *C. S. Rees et al.,* Theory and Applications of Fourier Analysis (1981)
60. *L. Nachbin,* Introduction to Functional Analysis (R. M. Aron, trans.) (1981)
61. *G. Orzech and M. Orzech,* Plane Algebraic Curves (1981)
62. *R. Johnsonbaugh and W. E. Pfaffenberger,* Foundations of Mathematical Analysis (1981)
63. *W. L. Voxman and R. H. Goetschel,* Advanced Calculus (1981)
64. *L. J. Corwin and R. H. Szczarba,* Multivariable Calculus (1982)
65. *V. I. Istrățescu,* Introduction to Linear Operator Theory (1981)
66. *R. D. Järvinen,* Finite and Infinite Dimensional Linear Spaces (1981)
67. *J. K. Beem and P. E. Ehrlich,* Global Lorentzian Geometry (1981)
68. *D. L. Armacost,* The Structure of Locally Compact Abelian Groups (1981)
69. *J. W. Brewer and M. K. Smith, eds.,* Emmy Noether: A Tribute (1981)
70. *K. H. Kim,* Boolean Matrix Theory and Applications (1982)
71. *T. W. Wieting,* The Mathematical Theory of Chromatic Plane Ornaments (1982)
72. *D. B. Gauld,* Differential Topology (1982)
73. *R. L. Faber,* Foundations of Euclidean and Non-Euclidean Geometry (1983)
74. *M. Carmeli,* Statistical Theory and Random Matrices (1983)
75. *J. H. Carruth et al.,* The Theory of Topological Semigroups (1983)
76. *R. L. Faber,* Differential Geometry and Relativity Theory (1983)
77. *S. Barnett,* Polynomials and Linear Control Systems (1983)
78. *G. Karpilovsky,* Commutative Group Algebras (1983)
79. *F. Van Oystaeyen and A. Verschoren,* Relative Invariants of Rings (1983)
80. *I. Vaisman,* A First Course in Differential Geometry (1984)
81. *G. W. Swan,* Applications of Optimal Control Theory in Biomedicine (1984)
82. *T. Petrie and J. D. Randall,* Transformation Groups on Manifolds (1984)
83. *K. Goebel and S. Reich,* Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings (1984)
84. *T. Albu and C. Năstăsescu,* Relative Finiteness in Module Theory (1984)
85. *K. Hrbacek and T. Jech,* Introduction to Set Theory: Second Edition (1984)
86. *F. Van Oystaeyen and A. Verschoren,* Relative Invariants of Rings (1984)
87. *B. R. McDonald,* Linear Algebra Over Commutative Rings (1984)
88. *M. Namba,* Geometry of Projective Algebraic Curves (1984)
89. *G. F. Webb,* Theory of Nonlinear Age-Dependent Population Dynamics (1985)
90. *M. R. Bremner et al.,* Tables of Dominant Weight Multiplicities for Representations of Simple Lie Algebras (1985)
91. *A. E. Fekete,* Real Linear Algebra (1985)
92. *S. B. Chae,* Holomorphy and Calculus in Normed Spaces (1985)
93. *A. J. Jerri,* Introduction to Integral Equations with Applications (1985)
94. *G. Karpilovsky,* Projective Representations of Finite Groups (1985)
95. *L. Narici and E. Beckenstein,* Topological Vector Spaces (1985)
96. *J. Weeks,* The Shape of Space (1985)
97. *P. R. Gribik and K. O. Kortanek,* Extremal Methods of Operations Research (1985)
98. *J.-A. Chao and W. A. Woyczynski, eds.,* Probability Theory and Harmonic Analysis (1986)
99. *G. D. Crown et al.,* Abstract Algebra (1986)
100. *J. H. Carruth et al.,* The Theory of Topological Semigroups, Volume 2 (1986)
101. *R. S. Doran and V. A. Belfi,* Characterizations of C*-Algebras (1986)
102. *M. W. Jeter,* Mathematical Programming (1986)
103. *M. Altman,* A Unified Theory of Nonlinear Operator and Evolution Equations with Applications (1986)
104. *A. Verschoren,* Relative Invariants of Sheaves (1987)
105. *R. A. Usmani,* Applied Linear Algebra (1987)
106. *P. Blass and J. Lang,* Zariski Surfaces and Differential Equations in Characteristic $p > 0$ (1987)
107. *J. A. Reneke et al.,* Structured Hereditary Systems (1987)
108. *H. Busemann and B. B. Phadke,* Spaces with Distinguished Geodesics (1987)
109. *R. Harte,* Invertibility and Singularity for Bounded Linear Operators (1988)

110. *G. S. Ladde et al.,* Oscillation Theory of Differential Equations with Deviating Arguments (1987)
111. *L. Dudkin et al.,* Iterative Aggregation Theory (1987)
112. *T. Okubo,* Differential Geometry (1987)
113. *D. L. Stancl and M. L. Stancl,* Real Analysis with Point-Set Topology (1987)
114. *T. C. Gard,* Introduction to Stochastic Differential Equations (1988)
115. *S. S. Abhyankar,* Enumerative Combinatorics of Young Tableaux (1988)
116. *H. Strade and R. Farnsteiner,* Modular Lie Algebras and Their Representations (1988)
117. *J. A. Huckaba,* Commutative Rings with Zero Divisors (1988)
118. *W. D. Wallis,* Combinatorial Designs (1988)
119. *W. Więsław,* Topological Fields (1988)
120. *G. Karpilovsky,* Field Theory (1988)
121. *S. Caenepeel and F. Van Oystaeyen,* Brauer Groups and the Cohomology of Graded Rings (1989)
122. *W. Kozlowski,* Modular Function Spaces (1988)
123. *E. Lowen-Colebunders,* Function Classes of Cauchy Continuous Maps (1989)
124. *M. Pavel,* Fundamentals of Pattern Recognition (1989)
125. *V. Lakshmikantham et al.,* Stability Analysis of Nonlinear Systems (1989)
126. *R. Sivaramakrishnan,* The Classical Theory of Arithmetic Functions (1989)
127. *N. A. Watson,* Parabolic Equations on an Infinite Strip (1989)
128. *K. J. Hastings,* Introduction to the Mathematics of Operations Research (1989)
129. *B. Fine,* Algebraic Theory of the Bianchi Groups (1989)
130. *D. N. Dikranjan et al.,* Topological Groups (1989)
131. *J. C. Morgan II,* Point Set Theory (1990)
132. *P. Biler and A. Witkowski,* Problems in Mathematical Analysis (1990)
133. *H. J. Sussmann,* Nonlinear Controllability and Optimal Control (1990)
134. *J.-P. Florens et al.,* Elements of Bayesian Statistics (1990)
135. *N. Shell,* Topological Fields and Near Valuations (1990)
136. *B. F. Doolin and C. F. Martin,* Introduction to Differential Geometry for Engineers (1990)
137. *S. S. Holland, Jr.,* Applied Analysis by the Hilbert Space Method (1990)
138. *J. Okniński,* Semigroup Algebras (1990)
139. *K. Zhu,* Operator Theory in Function Spaces (1990)
140. *G. B. Price,* An Introduction to Multicomplex Spaces and Functions (1991)
141. *R. B. Darst,* Introduction to Linear Programming (1991)
142. *P. L. Sachdev,* Nonlinear Ordinary Differential Equations and Their Applications (1991)
143. *T. Husain,* Orthogonal Schauder Bases (1991)
144. *J. Foran,* Fundamentals of Real Analysis (1991)
145. *W. C. Brown,* Matrices and Vector Spaces (1991)
146. *M. M. Rao and Z. D. Ren,* Theory of Orlicz Spaces (1991)
147. *J. S. Golan and T. Head,* Modules and the Structures of Rings (1991)
148. *C. Small,* Arithmetic of Finite Fields (1991)
149. *K. Yang,* Complex Algebraic Geometry (1991)
150. *D. G. Hoffman et al.,* Coding Theory (1991)
151. *M. O. González,* Classical Complex Analysis (1992)
152. *M. O. González,* Complex Analysis (1992)
153. *L. W. Baggett,* Functional Analysis (1992)
154. *M. Sniedovich,* Dynamic Programming (1992)
155. *R. P. Agarwal,* Difference Equations and Inequalities (1992)
156. *C. Brezinski,* Biorthogonality and Its Applications to Numerical Analysis (1992)
157. *C. Swartz,* An Introduction to Functional Analysis (1992)
158. *S. B. Nadler, Jr.,* Continuum Theory (1992)
159. *M. A. Al-Gwaiz,* Theory of Distributions (1992)
160. *E. Perry,* Geometry: Axiomatic Developments with Problem Solving (1992)
161. *E. Castillo and M. R. Ruiz-Cobo,* Functional Equations and Modelling in Science and Engineering (1992)
162. *A. J. Jerri,* Integral and Discrete Transforms with Applications and Error Analysis (1992)
163. *A. Charlier et al.,* Tensors and the Clifford Algebra (1992)
164. *P. Biler and T. Nadzieja,* Problems and Examples in Differential Equations (1992)
165. *E. Hansen,* Global Optimization Using Interval Analysis (1992)

166. *S. Guerre-Delabrière*, Classical Sequences in Banach Spaces (1992)
167. *Y. C. Wong*, Introductory Theory of Topological Vector Spaces (1992)
168. *S. H. Kulkarni and B. V. Limaye*, Real Function Algebras (1992)
169. *W. C. Brown*, Matrices Over Commutative Rings (1993)
170. *J. Loustau and M. Dillon*, Linear Geometry with Computer Graphics (1993)
171. *W. V. Petryshyn*, Approximation-Solvability of Nonlinear Functional and Differential Equations (1993)
172. *E. C. Young*, Vector and Tensor Analysis: Second Edition (1993)
173. *T. A. Bick*, Elementary Boundary Value Problems (1993)
174. *M. Pavel*, Fundamentals of Pattern Recognition: Second Edition (1993)
175. *S. A. Albeverio et al.*, Noncommutative Distributions (1993)
176. *W. Fulks*, Complex Variables (1993)
177. *M. M. Rao*, Conditional Measures and Applications (1993)
178. *A. Janicki and A. Weron*, Simulation and Chaotic Behavior of α-Stable Stochastic Processes (1994)
179. *P. Neittaanmäki and D. Tiba*, Optimal Control of Nonlinear Parabolic Systems (1994)
180. *J. Cronin*, Differential Equations: Introduction and Qualitative Theory, Second Edition (1994)
181. *S. Heikkilä and V. Lakshmikantham*, Monotone Iterative Techniques for Discontinuous Nonlinear Differential Equations (1994)
182. *X. Mao*, Exponential Stability of Stochastic Differential Equations (1994)
183. *B. S. Thomson*, Symmetric Properties of Real Functions (1994)
184. *J. E. Rubio*, Optimization and Nonstandard Analysis (1994)
185. *J. L. Bueso et al.*, Compatibility, Stability, and Sheaves (1995)
186. *A. N. Michel and K. Wang*, Qualitative Theory of Dynamical Systems (1995)
187. *M. R. Darnel*, Theory of Lattice-Ordered Groups (1995)
188. *Z. Naniewicz and P. D. Panagiotopoulos*, Mathematical Theory of Hemivariational Inequalities and Applications (1995)
189. *L. J. Corwin and R. H. Szczarba*, Calculus in Vector Spaces: Second Edition (1995)
190. *L. H. Erbe et al.*, Oscillation Theory for Functional Differential Equations (1995)
191. *S. Agaian et al.*, Binary Polynomial Transforms and Nonlinear Digital Filters (1995)
192. *M. I. Gil'*, Norm Estimations for Operation-Valued Functions and Applications (1995)
193. *P. A. Grillet*, Semigroups: An Introduction to the Structure Theory (1995)
194. *S. Kichenassamy*, Nonlinear Wave Equations (1996)
195. *V. F. Krotov*, Global Methods in Optimal Control Theory (1996)
196. *K. I. Beidar et al.*, Rings with Generalized Identities (1996)
197. *V. I. Arnautov et al.*, Introduction to the Theory of Topological Rings and Modules (1996)
198. *G. Sierksma*, Linear and Integer Programming (1996)
199. *R. Lasser*, Introduction to Fourier Series (1996)
200. *V. Sima*, Algorithms for Linear-Quadratic Optimization (1996)
201. *D. Redmond*, Number Theory (1996)
202. *J. K. Beem et al.* Global Lorentzian Geometry: Second Edition (1996)

Additional Volumes in Preparation

GLOBAL LORENTZIAN GEOMETRY

Second Edition

John K. Beem
Department of Mathematics
University of Missouri–Columbia
Columbia, Missouri

Paul E. Ehrlich
Department of Mathematics
University of Florida–Gainesville
Gainesville, Florida

Kevin L. Easley
Department of Mathematics
Truman State University
Kirksville, Missouri

Marcel Dekker, Inc.   New York • Basel • Hong Kon

Library of Congress Cataloging-in-Publication Data

Beem, John K.
Global Lorentzian geometry. — 2nd ed. / John K. Beem, Paul E. Ehrlich, Kevin L. Easley.
p. cm. — (Monographs and textbooks in pure and applied mathematics ; 202)
Includes bibliographical references and index.
ISBN 0-8247-9324-2 (pbk. : alk. paper)
1. Geometry, Differential. 2. General relativity (Physics).
I. Ehrlich, Paul E. II. Easley, Kevin L. III. Title. IV. Series.
QA649.B42 1996
516.3'7—dc20 96-957
CIP

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the address below.

This book is printed on acid-free paper.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE TO THE SECOND EDITION

The second edition of this book continues the study of Lorentzian geometry, the mathematical theory used in general relativity. Chapters 3 through 12 contain material, slightly revised in some cases, which was discussed in Chapters 2 through 11 of the first edition. Much new material has been added to Chapters 7 and 11, and new Chapters 13 and 14 have been written reflecting the more complete and detailed understanding that has been gained in the intervening years on many of the topics treated in the first edition. Inspired by an example of P. Williams (1984), additional material on the instability of both geodesic completeness and geodesic incompleteness for general space–times has been provided in Section 7.1. Section 7.4 has been added giving sufficient conditions on a space–time to guarantee the stability of geodesic completeness for metrics in a neighborhood of a given complete metric. New material has also been added to Section 11.3 on the topic of geodesic connectibility. Appendixes A, B, and C of the first edition have now been expanded into Chapter 2, which also contains new material on the generic condition as well as Section 2.3, which gives a proof that the null cone determines the metric up to a conformal factor in any semi-Riemannian manifold which is neither positive nor negative definite. Also, a deeper treatment of the behavior of the sectional curvature function in a neighborhood of a degenerate two-plane is given in Chapter 2.

In the concluding Chapter 11 of the first edition, which is now Chapter 12, we showed how the Lorentzian distance function and causally disconnecting sets could be used to obtain the Hawking–Penrose Singularity Theorem concerning geodesic incompleteness of the space–time manifold. Around 1980, S. T. Yau suggested that the concept of "curvature rigidity," well known for the differential geometry of complete Riemannian manifolds since the publication of Cheeger and Ebin (1975), might be applied to the seemingly unrelated topic of singularity theorems in space–time differential geometry. (Earlier, Geroch (1970b) had suggested that spatially closed space–times should fail to be timelike geodesically incomplete only under special circumstances.) As a step toward this conjectured rigidity of geodesic incompleteness, the Lorentzian analogue of the Cheeger–Gromoll Splitting Theorem for complete Riemannian manifolds of nonnegative Ricci curvature needed to be obtained. This was accomplished in a series of research papers published between 1984 and 1990. Aspects of the proof of the Lorentzian Splitting Theorem are discussed in the new Chapter 14. Another new chapter in the second edition, Chapter 13, draws upon investigations of Ehrlich and Emch (1992a,b,c, 1993) and is devoted to a study of the geodesic behavior and causal structure of a class of geodesically complete Ricci flat space–times, the gravitational plane waves, which were introduced into general relativity as astrophysical models. These space–times provide interesting and nontrivial examples of astigmatic conjugacy [cf. Penrose (1965a)] and of the nonspacelike cut locus, a concept discussed in Chapter 8 of the first edition.

As for the first edition, this book is written using the notations and methods of modern differential geometry. Thus the basic prerequisites remain a working knowledge of general topology and differential geometry. This book should be accessible to advanced graduate students in either mathematics or mathematical physics.

The list of works to which we are indebted for the two editions is quite extensive. In particular, we would like to single out the following five important sources: *The Large Scale Structure of Space–time* by S. W. Hawking and G. F. R. Ellis; *Techniques of Differential Topology in Relativity* by R. Penrose; *Riemannsche Geometrie im Grossen* by D. Gromoll, W. Klingenberg, and W. Meyer; the 1977 Diplomarbeit at Bonn University, *Existenz und Bedeutung von konjugierten Werten in der Raum–Zeit*, by G. Bölts; and *Semi-Riemannian Geometry* by B. O'Neill.

In the time from the late 1970's when we wrote the first edition to the present, it has been interesting to observe the enormous expansion in the journal literature on space–time differential geometry. This is reflected in the substantial growth of the list of references for the second edition. However, this wealth of new material has precluded our treating many interesting new developments in space–time geometry since 1980 which are less closely tied in with the overall viewpoint and selection of topics originally discussed in the first edition.

The authors would like to thank all those who have provided encouraging comments about the first edition and urged us to issue a second edition after the first edition had gone out of print, especially Gregory Galloway, Steven Harris, Andrzej Królak, Philip Parker, and Susan Scott. We thank Gerard Emch for insisting that the second edition be undertaken, and the first two authors thank Stephen Summers and Maria Allegra for independently suggesting that a third author be added to the team to share the duties of the completion of this revision. It is also a pleasure to thank Maria Allegra and Christine McCafferty at Marcel Dekker, Inc., for working with us to see the second edition into print. We are also indebted to Lia Petracovici for much helpful proofreading and to Todd Hammond for valuable technical advice concerning AMS-TEX.

John K. Beem
Paul E. Ehrlich
Kevin L. Easley

PREFACE TO THE FIRST EDITION

This book is about Lorentzian geometry, the mathematical theory used in general relativity, treated from the viewpoint of global differential geometry. Our goal is to help bridge the gap between modern differential geometry and the mathematical physics of general relativity by giving an invariant treatment of global Lorentzian geometry. The growing importance in physics of this approach is clearly illustrated by the recent Hawking–Penrose singularity theorems described in the text of Hawking and Ellis (1973).

The Lorentzian distance function is used as a unifying concept in our book. Furthermore, we frequently compare and contrast the results and techniques of Lorentzian geometry to those of Riemannian geometry to alert the reader to the basic differences between these two geometries.

This book has been written especially for the mathematician who has a basic acquaintance with Riemannian geometry and wishes to learn Lorentzian geometry. Accordingly, this book is written using the notation and methods of modern differential geometry. For readers less familiar with this notation, we have included Appendix A which gives the local coordinate representations for the symbols used.

The basic prerequisites for this book are a working knowledge of general topology and differential geometry. Thus this book should be accessible to advanced graduate students in either mathematics or mathematical physics.

In writing this monograph, both authors profited greatly from the opportunity to lecture on part of this material during the spring semester, 1978, at the University of Missouri–Columbia. The second author also gave a series of lectures on this material in Ernst Ruh's seminar in differential geometry at Bonn University during the summer semester, 1978, and would like to thank

Professor Ruh for giving him the opportunity to speak on this material. We would like to thank C. Ahlbrandt, D. Carlson, and M. Jacobs for several helpful conversations on Section 2.4 and the calculus of variations. We would like to thank M. Engman, S. Harris, K. Nomizu, T. Powell, D. Retzloff, and H. Wu for helpful comments on our preliminary version of this monograph. We also thank S. Harris for contributing Appendix D to this monograph and J.-H. Eschenburg for calling our attention to the Diplomarbeit of Bölts (1977). To anyone who has read either of the excellent books of Gromoll, Klingenberg, and Meyer (1975) on Riemannian manifolds or of Hawking and Ellis (1973) on general relativity, our debt to these authors in writing this work will be obvious. It is also a pleasure for both authors to thank the Research Council of the University of Missouri–Columbia and for the second author to thank the Sonderforschungsbereich Theoretische Mathematik 40 of the Mathematics Department, Bonn University, and to acknowledge an NSF Grant MCS77-18723(02) held at the Institute for Advanced Study, Princeton, New Jersey, for partial financial support while we were working on this monograph. Finally it is a pleasure to thank Diane Coffman, DeAnna Williamson, and Debra Retzloff for the patient and cheerful typing of the manuscript.

John K. Beem
Paul E. Ehrlich

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION: RIEMANNIAN THEMES IN LORENTZIAN GEOMETRY

In the 1970's, progress on causality theory, singularity theory, and black holes in general relativity, described in the influential text of Hawking and Ellis (1973), resulted in a resurgence of interest in global Lorentzian geometry. Indeed, a better understanding of global Lorentzian geometry was required for the development of singularity theory. For example, it was necessary to know that causally related points in globally hyperbolic subsets of space–times could be joined by a nonspacelike geodesic segment maximizing the Lorentzian arc length among all nonspacelike curves joining the two given points. In addition, much work done in the 1970's on foliating asymptotically flat Lorentzian manifolds by families of maximal hypersurfaces has been motivated by general relativity [cf. Choquet–Bruhat, Fischer, and Marsden (1979) for a partial list of references].

All of these results naturally suggested that a systematic study of global Lorentzian geometry should be made. The development of "modern" global Riemannian geometry as described in any of the standard texts [cf. Bishop and Crittenden (1964), Gromoll, Klingenberg, and Meyer (1975), Helgason (1978), Hicks (1965)] supported the idea that a comprehensive treatment of global Lorentzian geometry should be grounded in three fundamental topics: geodesic and metric completeness, the Lorentzian distance function, and a Morse index theory valid for nonspacelike geodesic segments in an arbitrary Lorentzian manifold.

Geodesic completeness, or more accurately geodesic incompleteness, played a crucial role in the development of singularity theory in general relativity and has been thoroughly explored within this framework. However, the Lorentzian distance function had not been as well investigated, although it had been of some use in the study of singularities [cf. Hawking (1967), Hawking and Ellis (1973), Tipler (1977a), Beem and Ehrlich (1979a)]. Some of the properties of the Lorentzian distance function needed in general relativity had been briefly described in Hawking and Ellis (1973, pp. 215–217). Further results relating Lorentzian distance to causality and the global behavior of nonspacelike geodesics had been given in Beem and Ehrlich (1979b). Hence, as part of the first edition, a systematic study of the Lorentzian distance function was made.

Uhlenbeck (1975), Everson and Talbot (1976), and Woodhouse (1976) had studied Morse index theory for globally hyperbolic space–times, and we had sketched [cf. Beem and Ehrlich (1979c,d)] a Morse index theory for nonspacelike geodesics in arbitrary space–times. But no complete treatment of this theory for arbitrary space–times had been published prior to the first edition of the current book.

The Lorentzian distance function has many similarities with the Riemannian distance function but also many differences. Since the Lorentzian distance function is still less well known, we now review the main properties of the Riemannian distance function and then compare and contrast the corresponding results for the Lorentzian distance function.

For the rest of this portion of the introduction, we will let (N, g_0) denote a Riemannian manifold and (M, g) denote a Lorentzian manifold, respectively.

Thus N is a smooth paracompact manifold equipped with a positive definite inner product $g_0|_p : T_pN \times T_pN \to \mathbb{R}$ on each tangent space T_pN. In addition, if X and Y are arbitrary smooth vector fields on N, the function $N \to \mathbb{R}$ given by $p \to g_0(X(p), Y(p))$ is required to be a smooth function. The Riemannian structure $g_0 : TN \times TN \to \mathbb{R}$ then defines the Riemannian distance function

$$d_0 : N \times N \to [0, \infty)$$

as follows. Let $\Omega_{p,q}$ denote the set of piecewise smooth curves in N from p to q. Given a curve $c \in \Omega_{p,q}$ with $c : [0,1] \to N$, there is a finite partition $0 = t_1 < t_2 < \cdots < t_k = 1$ such that $c|\,[t_i, t_{i+1}]$ is smooth for each i. The Riemannian arc length of c with respect to g_0 is defined as

$$L_0(c) = \sum_{i=1}^{k-1} \int_{t_i}^{t_{i+1}} \sqrt{g_0(c'(t), c'(t))}\, dt.$$

The Riemannian distance $d_0(p,q)$ between p and q is then defined to be

$$d_0(p,q) = \inf\{L_0(c) : c \in \Omega_{p,q}\} \geq 0.$$

For any Riemannian metric g_0 for N, the function $d_0 : N \times N \to [0,\infty)$ has the following properties:

(1) $d_0(p,q) = d_0(q,p)$ for all $p, q \in N$.
(2) $d_0(p,q) \leq d_0(p,r) + d_0(r,q)$ for all $p, q, r \in N$.
(3) $d_0(p,q) = 0$ if and only if $p = q$.

More surprisingly,

(4) $d_0 : N \times N \to [0,\infty)$ is continuous, and the family of metric balls

$$B(p,\epsilon) = \{q \in N : d_0(p,q) < \epsilon\}$$

for all $p \in N$ and $\epsilon > 0$ forms a basis for the given manifold topology.

Thus the metric topology and the given manifold topology coincide. Furthermore, by a result of Whitehead (1932), given any $p \in N$, there exists an $R > 0$ such that for any ϵ with $0 < \epsilon < R$, the metric ball $B(p,\epsilon)$ is geodesically convex. Thus for any ϵ with $0 < \epsilon < R$, the set $B(p,\epsilon)$ is diffeomorphic to the n-disk, $n = \dim(N)$, and the set $\{q \in N : d_0(p,q) = \epsilon\}$ is diffeomorphic to S^{n-1}.

Removing the origin from $\mathbb{R}^2$ equipped with the usual Euclidean metric and setting $p = (-1,0)$, $q = (1,0)$, one calculates that $d_0(p,q) = 2$ but finds no curve $c \in \Omega_{p,q}$ with $L_0(c) = d_0(p,q)$ and also no smooth geodesic from p to q. Thus the following questions arise naturally. Given a manifold N, find conditions on a Riemannian metric g_0 for N such that

(1) All geodesics in N may be extended to be defined on all of $\mathbb{R}$.
(2) The pair (N, d_0) is a complete metric space in the sense that all Cauchy sequences converge.
(3) Given any two points $p, q \in N$, there is a smooth geodesic segment $c \in \Omega_{p,q}$ with $L_0(c) = d_0(p,q)$.

A distance realizing geodesic segment as in (3) is called a *minimal* geodesic segment. The word *minimal* is used here since the definition of Riemannian distance implies that $L_0(\gamma) \geq d_0(p,q)$ for all $\gamma \in \Omega_{p,q}$. More generally, one may define an arbitrary piecewise smooth curve $\gamma \in \Omega_{p,q}$ to be *minimal* if $L_0(\gamma) = d_0(p,q)$. Using the variation theory of the arc length functional, it may be shown that if $\gamma \in \Omega_{p,q}$ is minimal, then γ may be reparametrized to a smooth geodesic segment.

The question of finding criteria on g_0 such that (1), (2), or (3) holds was resolved by H. Hopf and W. Rinow in their famous paper (1931). In modern terminology the Hopf–Rinow Theorem asserts the following.

Theorem (Hopf–Rinow). *For any Riemannian manifold (N, g_0) the following are equivalent:*

(1) *Metric completeness: (N, d_0) is a complete metric space.*

(2) *Geodesic completeness: For any $v \in TN$, the geodesic $c(t)$ in N with $c'(0) = v$ is defined for all positive and negative real numbers $t \in \mathbb{R}$.*

(3) *For some $p \in N$, the exponential map $\exp_p$ is defined on the entire tangent space T_pN to N at p.*

(4) *Finite compactness: Every subset K of N that is d_0 bounded (i.e., $\sup\{d_0(p,q) : p,q \in K\} < \infty$) has compact closure.*

Furthermore, if any one of (1) through (4) holds, then

(5) *Minimal geodesic connectibility: Given any $p, q \in N$, there exists a smooth geodesic segment c from p to q with $L_0(c) = d_0(p,q)$.*

A Riemannian manifold (N, g_0) is said to be *complete* provided any one (and hence all) of conditions (1) through (4) is satisfied. It should be stressed that the Hopf–Rinow Theorem guarantees the equivalence of metric and geodesic completeness and also that *all* Riemannian metrics for a compact smooth manifold are complete. Unfortunately, *none* of these statements is valid for arbitrary Lorentzian manifolds.

A remaining question for noncompact but paracompact manifolds is the existence of complete Riemannian metrics. This was settled by Nomizu and Ozeki's (1961) proof that given any Riemannian metric g_0 for N, there is a *complete* Riemannian metric for N globally conformal to g_0. Since any paracompact connected smooth manifold N admits a Riemannian metric by a partition of unity argument, it follows that N also admits a complete Riemannian metric.

We now turn our attention to the Lorentzian manifold (M, g). A Lorentzian metric g for the smooth paracompact manifold M is the assignment of a nondegenerate bilinear form $g|_p : T_pM \times T_pM \to \mathbb{R}$ with diagonal form $(-, +, \cdots, +)$ to each tangent space. It is well known that if M is compact and $\chi(M) \neq 0$, then M admits *no* Lorentzian metric. On the other hand, any noncompact manifold admits a Lorentzian metric. Geroch (1968a) and Marathe (1972) have also shown that a smooth Hausdorff manifold which admits a Lorentzian metric is paracompact.

Nonzero tangent vectors are classified as *timelike*, *spacelike*, *nonspacelike*, or *null* according to whether $g(v, v) < 0$, > 0, ≤ 0, or $= 0$, respectively. [Some authors use the convention $(+, -, \cdots, -)$ for the Lorentzian metric, and hence all of the inequality signs in the above definition are reversed for them.] A Lorentzian manifold (M, g) is said to be *time oriented* if M admits a continuous, nowhere vanishing, timelike vector field X. This vector field is used to separate the nonspacelike vectors at each point into two classes called *future directed* and *past directed*. A *space–time* is then a Lorentzian manifold (M, g) together with a choice of time orientation. We will usually work with space–times below.

In order to define the Lorentzian distance function and discuss its properties, we need to introduce some concepts from elementary causality theory. It is standard to write $p \ll q$ if there is a future directed piecewise smooth timelike curve in M from p to q, and $p \leqslant q$ if $p = q$ or if there is a future directed piecewise smooth nonspacelike curve in M from p to q. The *chronological past* and *future* of p are then given respectively by $I^-(p) = \{q \in M : q \ll p\}$ and $I^+(p) = \{q \in M : p \ll q\}$. The *causal past* and *future* of p are defined as $J^-(p) = \{q \in M : q \leqslant p\}$ and $J^+(p) = \{q \in M : p \leqslant q\}$. The sets $I^-(p)$ and $I^+(p)$ are always open in any space–time, but the sets $J^-(p)$ and $J^+(p)$ are neither open nor closed in general (cf. Figure 1.1).

The *causal structure* of the space–time (M, g) may be defined as the collection of past and future sets at all points of M together with their properties.

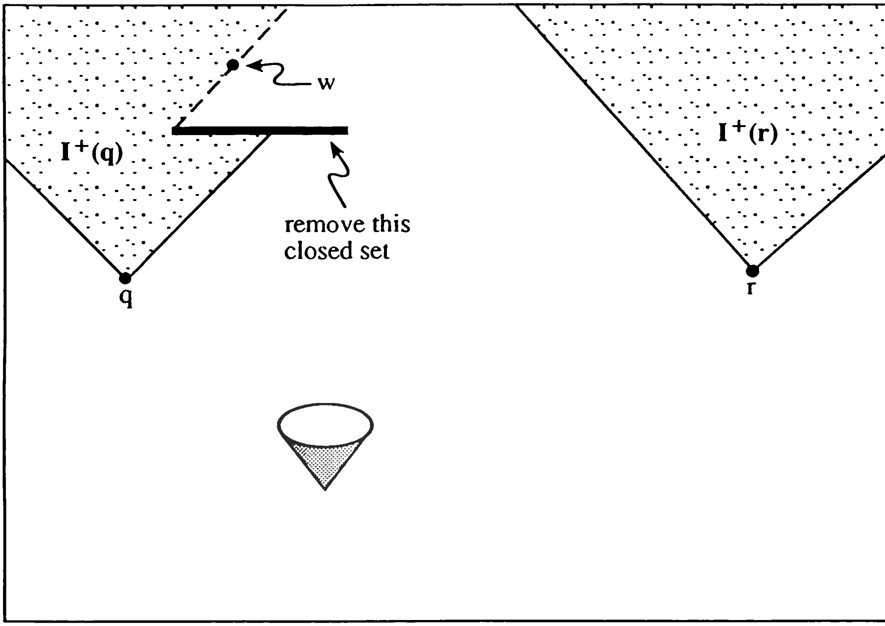


FIGURE 1.1. The chronological (respectively, causal) future of a point consists of all points which can be reached from that point by future directed timelike (respectively, nonspacelike) curves. In this example, the causal future $J^+(r)$ of r is the closure of the chronological future $I^+(r)$ of r. However, the set $J^+(q)$ is not the closure of $I^+(q)$. In particular, the point w is in the closure of $I^+(q)$ but is not in $J^+(q)$.

It may be shown that two strongly causal Lorentzian metrics g_1 and g_2 for M determine the same past and future sets at all points if and only if the two metrics are globally conformal [i.e., $g_1 = \Omega g_2$ for some smooth function $\Omega : M \to (0, \infty)$]. Letting $C(M, g)$ denote the set of Lorentzian metrics globally conformal to g, it follows that properties suitably defined using the past and future sets hold simultaneously either for all metrics in $C(M, g)$ or for no metric in $C(M, g)$. Thus all of the basic properties of elementary causality theory depend only on the conformal class $C(M, g)$ and not on the choice of Lorentzian metric representing $C(M, g)$.

Perhaps the two most elementary properties to require of the conformal structure $C(M, g)$ are either that (M, g) be chronological or that (M, g) be causal. A space–time (M, g) is said to be *chronological* if $p \notin I^+(p)$ for all $p \in M$. This means that (M, g) contains no closed timelike curves. The space–time (M, g) is said to be *causal* if there is no pair of distinct points $p, q \in M$ with $p \leqslant q \leqslant p$. This is equivalent to requiring that (M, g) contain no closed nonspacelike curves.

Already at this stage, a basic difference emerges between Lorentzian and Riemannian geometry. On physical grounds, the space–times of general relativity are usually assumed to be chronological. But it is easy to show that if M is compact, (M, g) contains a closed timelike curve. Thus the space–times usually considered in general relativity are assumed to be noncompact.

In general relativity each point of a Lorentzian manifold corresponds to an event. Thus the existence of a closed timelike curve raises the possibility that a person might traverse some path and meet himself at an earlier age. More generally, closed nonspacelike curves generate paradoxes involving causality and are thus said to "violate causality." Even if a space–time has no closed nonspacelike curves, it may contain a point p such that there are future directed nonspacelike curves leaving arbitrarily small neighborhoods of p and then returning. This behavior is said to be a violation of strong causality at p. Space–times with no such violation are *strongly causal.*

The strongly causal space–times form an important subclass of the causal space–times. For this class of space–times the Alexandrov topology with basis $\{I^+(p) \cap I^-(q) : p, q \in M\}$ for M and the given manifold topology are related as follows [cf. Kronheimer and Penrose (1967), Penrose (1972)].

Theorem. *The following are equivalent:*

(1) *(M, g) is strongly causal.*

(2) *The Alexandrov topology induced on M agrees with the given manifold topology.*

(3) *The Alexandrov topology is Hausdorff.*

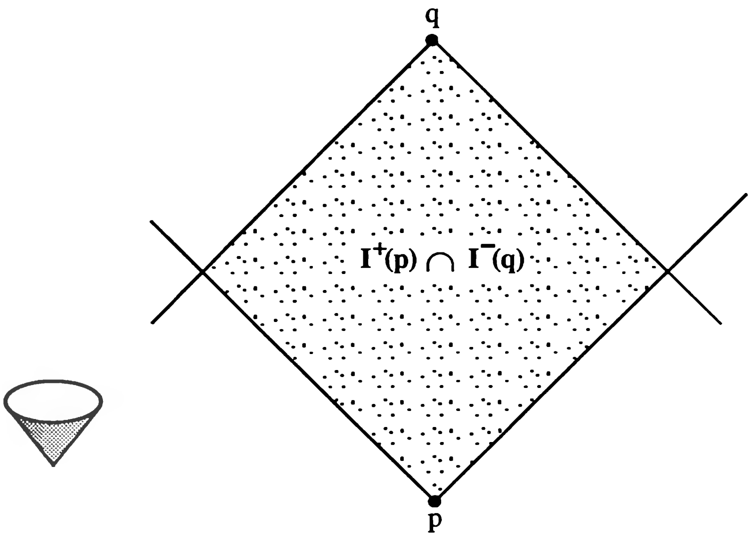


FIGURE 1.2. Sets of the form $I^+(p) \cap I^-(q)$ with arbitrary $p, q \in M$ form a basis of the Alexandrov topology. This topology is always at least as coarse as the original topology on M. The Alexandrov topology agrees with the original topology if and only if (M, g) is strongly causal.

We are ready at last to define the *Lorentzian distance function*

$$d = d(g) : M \times M \to [0, \infty]$$

of an arbitrary space–time. If $c : [0, 1] \to M$ is a piecewise smooth nonspacelike curve differentiable except at $0 = t_1 < t_2 < \cdots < t_k = 1$, then the length $L(c) = L_g(c)$ of c is given by the formula

$$L(c) = \sum_{i=1}^{k-1} \int_{t_i}^{t_{i+1}} \sqrt{-g(c'(t), c'(t))}\, dt.$$

If $p \ll q$, there are timelike curves from p to q (very close to piecewise null curves) of arbitrarily small length. Hence the infimum of Lorentzian arc length of all piecewise smooth curves joining any two chronologically related points

$p \ll q$ is zero. On the other hand, if $p \ll q$ and p and q lie in a geodesically convex neighborhood U, the future directed timelike geodesic segment in U from p to q has the largest Lorentzian arc length among all nonspacelike curves in U from p to q. Thus the following definition for $d(p, q)$ is natural: fixing a point $p \in M$, set $d(p, q) = 0$ if $q \notin J^+(p)$, and otherwise calculate $d(p, q)$ for $q \in J^+(p)$ as the supremum of Lorentzian arc length of all future directed nonspacelike curves from p to q. Thus if $q \in J^+(p)$ and γ is any future directed nonspacelike curve from p to q, we have $L(\gamma) \leq d(p, q)$. Hence unlike the Riemannian distance function, the Lorentzian distance function is not a priori finite-valued. Indeed, a so-called totally vicious space–time may be characterized in terms of its Lorentzian distance function by the property that $d(p, q) = \infty$ for all $p, q \in M$. Also, if (M, g) is nonchronological and $p \in I^+(p)$, it follows that $d(p, p) = \infty$.

The Reissner–Nordström space–times, physically important examples of exact solutions to the Einstein equations in general relativity, also contain pairs of chronologically related distinct points $p \ll q$ with $d(p, q) = \infty$.

By definition of Lorentzian distance, $d(p, q) = 0$ whenever $q \in M - J^+(p)$. We have even seen that $d(p, p) = \infty$ is possible. Thus for arbitrary Lorentzian manifolds there is no analogue for property (3) of the Riemannian distance function. Also, the Lorentzian distance function tends from its definition to be a nonsymmetric distance. In particular, for any space–time it may be shown that if $0 < d(p, q) < \infty$, then $d(q, p) = 0$. But the Lorentzian distance function does possess a useful analogue for property (2) of the Riemannian distance function. Namely, $d(p, q) \geq d(p, r) + d(r, q)$ for all $p, q, r \in M$ with $p \leqslant r \leqslant q$. The reversal of inequality sign is to be expected since nonspacelike geodesics in a Lorentzian manifold locally maximize rather than minimize arc length.

Since $d(p, q) > 0$ if and only if $q \in I^+(p)$, and $d(q, p) > 0$ if and only if $q \in I^-(p)$, the distance function determines the chronology of (M, g). On the other hand, conformally changing the metric changes distance but not the chronology, so that the chronology does not determine the distance function. Clearly, the distance function does not determine the sets $J^+(p)$ or $J^-(p)$ since $d(p, q) = 0$ not only for $q \in J^+(p) - I^+(p)$ but also for $q \in M - J^+(p)$.

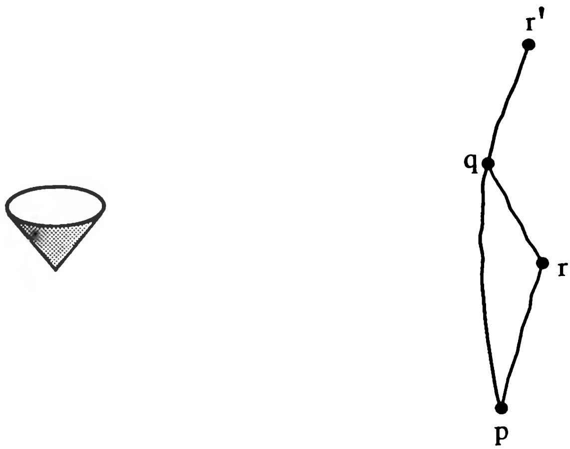


FIGURE 1.3. If r is in the causal future of p and q is in the causal future of r, then the distance function satisfies the reverse triangle inequality $d(p,q) \geq d(p,r) + d(r,q)$. The reverse triangle inequality will not be valid in general for some point r' which is not causally between p and q.

Property (4) of the Riemannian distance function is the continuity of this function for all Riemannian metrics. For space–times, on the other hand, the Lorentzian distance function may fail to be upper semicontinuous. Indeed, the continuity of $d : M \times M \to [0, \infty]$ has the following consequence for the causal structure of (M, g) [cf. Theorem 4.24]. If (M, g) is a distinguishing space–time and d is continuous, then (M, g) is causally continuous (cf. Chapter 3 for definitions of these concepts). Hence it is necessary to accept the lack of continuity and lack of finiteness of the Lorentzian distance function for arbitrary space–times. The Lorentzian distance function is at least lower semicontinuous where it is finite. This may be combined with the upper semicontinuity in the C^0 topology of the Lorentzian arc length functional for strongly causal space–times to construct distance realizing nonspacelike geodesics in certain classes of space–times (cf. Sections 8.1 and 8.2).

With these remarks in mind, it is natural to ask if some class of space–times for which the Lorentzian distance function is finite-valued and/or continuous may be found. The globally hyperbolic space–times turn out to be such a class. Here a space–time (M, g) is said to be *globally hyperbolic* if it is strongly causal and satisfies the condition that $J^+(p) \cap J^-(q)$ is compact for all $p, q \in M$. It has been most useful in proving singularity theorems in general relativity to know that if (M, g) is globally hyperbolic, then its Lorentzian distance function is finite-valued and continuous. Oddly enough, the finiteness of the distance function, rather than its continuity, characterizes globally hyperbolic space–times in the following sense (cf. Theorem 4.30). We say that a space–time (M, g) satisfies the *finite distance condition* provided that $d(g)(p, q) < \infty$ for all $p, q \in M$. It may then be shown that the strongly causal Lorentzian manifold (M, g) is globally hyperbolic if and only if (M, g') satisfies the finite distance condition for all $g' \in C(M, g)$.

Motivated by the definition of minimal curve in Riemannian geometry, we make the following definition for space–times.

Definition. (*Maximal Curve*) A future directed nonspacelike curve γ from p to q is said to be *maximal* if $L(\gamma) = d(p, q)$.

It may be shown (cf. Theorem 4.13), just as for minimal curves in Riemannian spaces, that if γ is a maximal curve from p to q, then γ may be reparametrized to a nonspacelike geodesic segment. This result may be used to construct geodesics in strongly causal space–times as limit curves of appropriate sequences of "almost maximal" nonspacelike curves (cf. Sections 8.1, 8.2).

In view of (5) of the Hopf–Rinow Theorem for Riemannian manifolds, it is reasonable to look for a class of space–times satisfying the property that if $p \leqslant q$, there is a maximal geodesic segment from p to q. Using the compactness of $J^+(p) \cap J^-(q)$, one can show that globally hyperbolic space–times always contain maximal geodesics joining any two causally related points.

We are finally led to consider what can be said about Lorentzian analogues for the rest of the Hopf–Rinow Theorem. Here, however, every conceivable thing goes wrong. Thus much of the difficulty in Lorentzian geometry from the viewpoint of global Riemannian geometry or its richness from the viewpoint of singularity theory in general relativity stems from the lack of a sufficiently strong analogue of the Hopf–Rinow Theorem.

We now give a basic definition which corresponds to (2) of the Hopf–Rinow Theorem.

Definition. (*Timelike, Null, and Spacelike Geodesic Completeness*) A space–time (M, g) is said to be *timelike* (respectively, *null, spacelike, nonspacelike*) *complete* if all timelike (respectively, null, spacelike, nonspacelike) geodesics may be defined for all values $-\infty < t < \infty$ of an affine parameter t.

A space–time which is nonspacelike incomplete thus has a timelike or null geodesic which cannot be defined for all values of an affine parameter. Such space–times are said to be *singular* in the theory of general relativity.

It is first important to note that global hyperbolicity does not imply any of these forms of geodesic completeness. This may be seen by fixing points p and q in Minkowski space with $p \ll q$ and equipping $M = I^+(p) \cap I^-(q)$ with the Lorentzian metric it inherits as an open subset of Minkowski space. This space–time M is globally hyperbolic. Since geodesics in M are just the restriction of geodesics in Minkowski space to M, it follows that *every* geodesic in M is incomplete.

It was once hoped that timelike completeness might imply null completeness, etc. However, a series of examples has been given by Kundt, Geroch, and Beem of globally hyperbolic space–times for which timelike geodesic completeness, null geodesic completeness, and spacelike geodesic completeness are all logically inequivalent. Thus, there are globally hyperbolic space–times that are spacelike and timelike complete but null incomplete.

Metric completeness and geodesic completeness [(1) iff (2) of the Hopf–Rinow Theorem] are unrelated for arbitrary Lorentzian manifolds. There are also Lorentzian metrics which are timelike geodesically complete but also contain points p and q with $p \ll q$ such that *no* timelike geodesic from p to q exists (cf. Figure 6.1).

On the brighter side, there is some relationship between (1) and (4) of the Hopf–Rinow Theorem for globally hyperbolic space–times. Since $d(p, q) = 0$

if $q \notin J^+(p)$, convergence of arbitrary sequences in (M, g) with respect to the Lorentzian distance function does not make sense. But timelike Cauchy completeness may be defined (cf. Section 6.3). It can be shown for globally hyperbolic space–times that timelike Cauchy completeness and a type of finite compactness are equivalent.

In addition, results analogous to the Nomizu–Ozeki Theorem mentioned above for Riemannian metrics have been obtained. For instance, given any strongly causal space–time (M, g), there is a conformal factor $\Omega : M \to (0, \infty)$ such that the space–time $(M, \Omega g)$ is timelike and null geodesically complete (cf. Theorem 6.5). It is still unknown, however, whether this result can be strengthened to include spacelike geodesic completeness as well.

It should now be clear that while there are certain similarities between the Lorentzian and the Riemannian distance functions, especially for globally hyperbolic space–times, there are also striking differences. In spite of these differences, the Lorentzian distance function has many uses similar to those of the Riemannian distance function.

In Chapter 8 the Lorentzian distance function is used in constructing maximal nonspacelike geodesics. These maximal geodesics play a key role in the proof of singularity theorems (cf. Chapter 12). In Chapter 9 the Lorentzian distance function is used to define and study the Lorentzian cut locus.

In Chapter 10 a Morse index theory is developed for both timelike and null geodesics. A number of global results for Lorentzian manifolds are obtained in Chapter 11 using the index theory and the Lorentzian distance function. Also, results are presented concerning the relationship of geodesic connectibility to geodesic pseudoconvexity and geodesic disprisonment. In Chapter 13 a nontrivial example is given of the cut locus structure and certain associated achronal boundaries for the class of gravitational plane wave space–times from general relativity. Finally, Chapter 14 treats the concept of rigidity of geodesic incompleteness and the Lorentzian Splitting Theorem for space–times satisfying the timelike convergence condition which also contain a complete Lorentzian distance-realizing timelike geodesic line. In this setting, the almost maximal curves of Chapter 8 again play a role [cf. Galloway and Horta (1995)].

CHAPTER 2

CONNECTIONS AND CURVATURE

Let (M, g) be an n-dimensional manifold M with a semi-Riemannian metric g of arbitrary signature $(-, \cdots, -, +, \cdots, +)$. Then there exists a unique connection ∇ on M which is both compatible with the metric tensor g and torsion free. This connection, which is called the *Levi–Civita connection* of (M, g), satisfies the same formal relations whether or not (M, g) is positive definite. Thus geodesics, curvature, Ricci curvature, scalar curvature, and sectional curvature may all be defined for semi-Riemannian manifolds using the same formulas as for positive definite Riemannian manifolds. The only difficulty which arises is that the sectional curvature is not defined for degenerate sections of the tangent space when (M, g) is not of constant curvature. In fact, the sectional curvature is only bounded near all degenerate sections at a point in the case of constant sectional curvature [see Kulkarni (1979)]. This generic "blow up" of the sectional curvature at degenerate sections corresponds to a generic unboundedness of tidal accelerations [cf. Beem and Parker (1990)] for objects moving arbitrarily close to the speed of light.

In the first section of this chapter there is an introduction to connections, and in the second section semi-Riemannian manifolds are discussed. Riemannian manifolds are positive definite semi-Riemannian manifolds and thus have metrics of signature $(+, +, \cdots, +)$. Consequently, the metric induced on the tangent space of a Riemannian manifold is Euclidean. The types of semi-Riemannian manifolds of primary interest in this book are Lorentzian manifolds. These manifolds have metric tensors of signature $(-, +, \cdots, +)$. Thus, the tangent spaces of a Lorentzian manifold have induced Minkowskian metrics. The tangent vectors at $p \in M$ of length zero form the null cone at p. For semi-Riemannian manifolds which are neither positive nor negative definite, a null cone is an $(n-1)$-dimensional surface in the tangent space. In the third section of this chapter we show that null cones determine the metric up to a conformal factor for metrics which are not definite. In the fourth section we consider sectional curvature. This curvature is related to tidal accelerations using the Jacobi equation. The Jacobi equation analyzes the relative behavior of "nearby" geodesics and will be of fundamental importance in later chapters. We introduce the generic condition in the fifth section. This condition corresponds to a tidal acceleration assumption. The Einstein equations are introduced in the last section. These are partial differential equations relating the metric tensor and its first two derivatives to the energy-momentum tensor T. The Einstein equations thus link geometry in terms of the metric and curvature to physics in terms of the distribution of mass and energy.

The manifold M will always be smooth, paracompact, and Hausdorff. The tangent bundle will be denoted by TM, and the tangent space at $p \in M$ will be denoted by T_pM. If (U, x) is an arbitrary chart for M, then $(x^1, x^2, \ldots, x^n)$ will denote local coordinates on M and $\partial/\partial x^1, \partial/\partial x^2, \ldots, \partial/\partial x^n$ will denote the natural basis for the tangent space.

2.1 Connections

Let $\mathfrak{X}(M)$ denote the set of all smooth vector fields defined on M, and let $\mathfrak{F}(M)$ denote the ring of all smooth real-valued functions on M. A *connection* is a function

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$$

with the properties that

$$\nabla_V(X+Y) = \nabla_V X + \nabla_V Y, \tag{2.1}$$

$$\nabla_{fV+hW}(X) = f\nabla_V X + h\nabla_W X, \quad \text{and} \tag{2.2}$$

$$\nabla_V(fX) = f\nabla_V X + V(f)X \tag{2.3}$$

for all $f, h \in \mathfrak{F}(M)$ and all $X, Y, V, W \in \mathfrak{X}(M)$.

The vector $\nabla_{X(p)}Y = \nabla_X Y|_p$ at the point $p \in M$ depends only on the connection ∇, the value $X(p) = X_p$ of X at p, and the values of Y along any smooth curve which passes through p and has tangent $X(p)$ at p [cf. Hicks (1965, p. 57)]. To see this, let $E_1, E_2, \ldots, E_n$ be smooth vector fields defined near p which form a basis of the tangent space at each point in a neighborhood of p. Then $X(p) = \sum X^i(p)E_i(p)$ and $Y = \sum Y^i E_i$. Hence

$$\begin{aligned}\nabla_X Y|_p &= \nabla_{X(p)}\left(\sum_{i=1}^{n} Y^i E_i\right)\\ &= \sum_{i=1}^{n} Y^i(p)\nabla_{X(p)}E_i + \sum_{i=1}^{n} X(p)(Y^i)E_i(p)\\ &= \sum_{i,j=1}^{n} X^j(p)Y^i(p)\nabla_{E_j(p)}E_i + \sum_{i=1}^{n} X(p)(Y^i)E_i(p).\end{aligned}$$

It follows that $X^i(p)$, $Y^i(p)$, and $X(p)(Y^i)$ determine $\nabla_X Y|_p$ completely if the $\nabla_{E_j(p)}E_i$'s are known.

Given the connection ∇ on M and a curve $\gamma : [a,b] \to M$, we may define parallel translation of vector fields along γ. Here a *vector field Y along γ* is a smooth mapping $Y : [a,b] \to TM$ such that $Y(t) \in T_{\gamma(t)}M$ for each $t \in [a,b]$. For $t_0 \in [a,b]$ we may locally extend Y to a smooth vector field defined on a neighborhood of $\gamma(t_0)$. Then we may consider the vector field $\nabla_{\gamma'(t)}Y$ along γ. The preceding arguments show that this vector field along γ is independent of the local extension, and consequently $\nabla_{\gamma'}Y\ (= Y')$ is well defined. A vector field Y along γ which satisfies $\nabla_{\gamma'}Y(t) = 0$ for all $t \in [a,b]$ is said to move by *parallel translation* along γ. A *geodesic* $c : (a,b) \to M$ is a smooth curve of M such that the tangent vector c' moves by parallel translation along c. In other words, c is a geodesic if

$$\nabla_{c'}c' = 0. \qquad (\textit{geodesic equation}) \tag{2.4}$$

A *pregeodesic* is a smooth curve c which may be reparametrized to be a geodesic. Any parameter for which c is a geodesic is called an *affine parameter*. If s and t are two affine parameters for the same pregeodesic, then $s = at + b$ for some constants $a, b \in \mathbb{R}$. A pregeodesic is said to be *complete* if for some affine parametrization (hence for all affine parametrizations) the domain of the parametrization is all of $\mathbb{R}$. The equation $\nabla_{c'}c' = 0$ may be expressed as a system of linear differential equations. To this end, we let $(U, (x^1, x^2, \dots, x^n))$ be local coordinates on M and let $\partial/\partial x^1, \partial/\partial x^2, \dots, \partial/\partial x^n$ denote the *natural basis* with respect to these coordinates. The *connection coefficients* Γ^k_{ij} of ∇ with respect to $(x^1, x^2, \dots, x^n)$ are defined by

$$\nabla_{\partial/\partial x^i}\left(\frac{\partial}{\partial x^j}\right) = \sum_{k=1}^{n} \Gamma^k_{ij} \frac{\partial}{\partial x^k}. \tag{2.5}$$ (*connection coefficients*)

Using these coefficients we may write equation (2.4) as the system

$$\frac{d^2x^k}{dt^2} + \sum_{i,j=1}^{n} \Gamma^k_{ij} \frac{dx^i}{dt}\frac{dx^j}{dt} = 0. \tag{2.6}$$ (*geodesic equations in coordinates*)

The connection coefficients (*Christoffel symbols of the second kind*) may also be used to give a local representation of the action of ∇. If the vector fields X and Y have local representations as

$$X = \sum_{i=1}^{n} X^i(x)\frac{\partial}{\partial x^i} \qquad \text{and} \qquad Y = \sum_{i=1}^{n} Y^i(x)\frac{\partial}{\partial x^i},$$

then $\nabla_X Y$ has a local representation

$$\nabla_X Y = \sum_{k=1}^{n}\left(\sum_{j=1}^{n} X^j \frac{\partial Y^k}{\partial x^j} + \sum_{i,j=1}^{n} \Gamma^k_{ji} X^j Y^i\right)\frac{\partial}{\partial x^k}. \tag{2.7}$$

The vector field $\nabla_X Y$ is said to be the *covariant derivative* of Y with respect to X. A semicolon is used to denote covariant differentiation with respect to a natural basis vector. If $X = \partial/\partial x^j$, then the components of $\nabla_X Y = \nabla_{\partial/\partial x^j} Y$ are denoted by $Y^k{}_{;j}$ where

$$Y^k{}_{;j} = \frac{\partial Y^k}{\partial x^j} + \sum_{i=1}^{n} \Gamma^k_{ji} Y^i.$$

The *Lie bracket* of the ordered pair of vector fields X and Y is a vector field $[X, Y]$ which acts on a smooth function f by $[X, Y](f) = X(Y(f)) - Y(X(f))$. If $X = \sum X^i(x)\partial/\partial x^i$ and $Y = \sum Y^i(x)\partial/\partial x^i$, then

$$[X, Y] = \sum_{i,j=1}^{n}\left(X^i \frac{\partial Y^j}{\partial x^i} - Y^i \frac{\partial X^j}{\partial x^i}\right)\frac{\partial}{\partial x^j}.$$ (*Lie bracket*)

The *torsion tensor* T of ∇ is the function $T : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ given by

$$(2.8) \qquad T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]. \qquad (\textit{torsion tensor})$$

The mapping T is said to be *f-bilinear* since it is linear in both arguments and also satisfies $T(fX,Y) = T(X,fY) = fT(X,Y)$ for smooth functions f. The value $T(X,Y)|_p$ depends only on the connection ∇ and the values $X(p)$ and $Y(p)$. Consequently, T determines a bilinear map $T_pM \times T_pM \to T_pM$ at each point $p \in M$. Using the skew symmetry ($[X,Y] = -[Y,X]$) of the Lie bracket, it is easily seen that $T(X,Y) = -T(Y,X)$, and hence T is also skew. Since $[\partial/\partial x^i, \partial/\partial x^j] = 0$ for all $1 \le i,j \le n$, it follows that

$$(2.9) \qquad T\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) = \sum_{k=1}^{n} \left(\Gamma^k_{ij} - \Gamma^k_{ji}\right) \frac{\partial}{\partial x^k}.$$

Consequently,

$$T = \sum_{i,j,k=1}^{n} T_i{}^k{}_j \, dx^i \otimes \frac{\partial}{\partial x^k} \otimes dx^j$$

where

$$T_i{}^k{}_j = \Gamma^k_{ij} - \Gamma^k_{ji}. \qquad (\textit{torsion components})$$

Clearly, the torsion tensor provides a measure of the nonsymmetry of the connection coefficients. Hence, $T = 0$ if and only if these coefficients are symmetric in their subscripts. A connection ∇ with $T = 0$ is said to be *torsion free* or *symmetric*.

The *curvature* R of ∇ is a function which assigns to each pair $X, Y \in \mathfrak{X}(M)$ the f-linear map $R(X,Y) : \mathfrak{X}(M) \to \mathfrak{X}(M)$ given by

$$(2.10) \qquad R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \qquad (\textit{curvature})$$

Curvature provides a measure of the noncommutativity of ∇_X and ∇_Y. It should be noted that some authors define the curvature as the negative of the above definition. Consequently, they differ in sign for some of the definitions of curvature quantities given below. The curvature R represents a tensor field.

Hence, the map $(X, Y, Z) \to R(X, Y)Z$ from $\mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M)$ to $\mathfrak{X}(M)$ is f-trilinear, and the vector $R(X, Y)Z|_p$ depends only on $X(p)$, $Y(p)$, $Z(p)$, and ∇. If $x, y, z \in T_pM$, one may extend these vectors to corresponding smooth vector fields X, Y, Z and define $R(x, y)z = R(X, Y)Z|_p$.

If $\omega \in T_p^*M$ is a cotangent vector at p and $x, y, z \in T_pM$ are tangent vectors at p, then one defines

$$R(\omega, x, y, z) = (\omega, R(X, Y)Z) = \omega(R(X, Y)Z) \tag{2.11}$$

for X, Y, and Z smooth vector fields extending x, y, and z, respectively. The *curvature tensor* R is a $(1, 3)$ tensor field which is given in local coordinates by

$$R = \sum_{i,j,k,m=1}^{n} R^i{}_{jkm} \frac{\partial}{\partial x^i} \otimes dx^j \otimes dx^k \otimes dx^m \tag{2.12}$$

where the *curvature components* $R^i{}_{jkm}$ are given by

$$R^i{}_{jkm} = \frac{\partial \Gamma^i_{mj}}{\partial x^k} - \frac{\partial \Gamma^i_{kj}}{\partial x^m} + \sum_{a=1}^{n} (\Gamma^a_{mj} \Gamma^i_{ka} - \Gamma^a_{kj} \Gamma^i_{ma}). \tag{2.13}$$

Notice that $R(X, Y)Z = -R(Y, X)Z$, $R(\omega, X, Y, Z) = -R(\omega, Y, X, Z)$, and $R^i{}_{jkm} = -R^i{}_{jmk}$. Furthermore, if $X = \sum X^i \partial/\partial x^i$, $Y = \sum Y^i \partial/\partial x^i$, $Z = \sum Z^i \partial/\partial x^i$, and $\omega = \sum \omega_i dx^i$, then

$$R(X, Y)Z = \sum_{i,j,k,m=1}^{n} R^i{}_{jkm} Z^j X^k Y^m \frac{\partial}{\partial x^i} \tag{2.14}$$

and

$$R(\omega, X, Y, Z) = \sum_{i,j,k,m=1}^{n} R^i{}_{jkm} \omega_i Z^j X^k Y^m. \tag{2.15}$$

Consequently, one has $R(dx^i, \partial/\partial x^k, \partial/\partial x^m, \partial/\partial x^j) = R^i{}_{jkm}$.

2.2 Semi-Riemannian Manifolds

A *semi-Riemannian metric* g for a manifold M is a smooth symmetric tensor field of type $(0, 2)$ on M which assigns to each point $p \in M$ a nondegenerate inner product $g|_p : T_pM \times T_pM \to \mathbb{R}$ of signature $(-, \cdots, -, +, \cdots, +)$. Here *nondegenerate* means that for each nontrivial vector $v \in T_pM$ there is some $w \in T_pM$ such that $g_p(v, w) \neq 0$. If g has components g_{ij} in local coordinates, then the nondegeneracy assumption is equivalent to the condition that the determinant of the matrix (g_{ij}) be nonzero.

Although we consider only smooth metrics, some authors have studied distributional semi-Riemannian metrics [cf. Parker (1979), Taub (1980)]. Also, a number of authors have studied semi-Riemannian metrics for which degeneracy is allowed [cf. Bejancu and Duggal (1991), Katsuno (1980), Kossowski (1987, 1989)].

In local coordinates $(U, (x^1, x^2, \ldots, x^n))$ on M, the metric g is represented by

$$g \,|\, U = \sum_{i,j=1}^{n} g_{ij}(x)\, dx^i \otimes dx^j \qquad (\textit{metric tensor})$$

with

$$g_{ij} = g_{ji} \qquad (\textit{symmetric}) \qquad \text{and} \qquad \det(g_{ij}) \neq 0. \qquad (\textit{nondegenerate})$$

If g has s negative eigenvalues and $r = n - s$ positive eigenvalues, then the *signature* of g will be denoted by (s, r). For each fixed $p \in M$, there exist local coordinates $(U, (x^1, x^2, \ldots, x^n))$ such that $g_p = g \,|\, T_pM$ can be represented as the diagonal matrix $\text{diag}\{-, \cdots, -, +, \cdots, +\}$. For each semi-Riemannian manifold (M, g) there is an associated semi-Riemannian manifold $(M, -g)$ obtained by replacing g with $-g$. Aside from some minor changes in sign, there is no essential difference between (M, g) and $(M, -g)$. Thus, results for spaces of signature (s, r) may always be translated into corresponding results for spaces of signature (r, s) by appropriate sign changes and inequality reversals.

Two vectors in T_pM are *orthogonal* if their inner product with respect to g_p is zero. Note that when g has eigenvalues with different signs, there will be some nontrivial vectors which are orthogonal to themselves (i.e., satisfy $g(v, v) = 0$). These are known as *null vectors*. A given vector is said to be a *unit vector* if it has inner product with itself equal to either $+1$ or -1. Thus, an *orthonormal basis* $\{e_1, e_2, \ldots, e_n\}$ of T_pM satisfies $|g(e_i, e_j)| = \delta^i{}_j$.

Given a semi-Riemannian manifold (M, g), there is a unique connection ∇ on M such that

$$(2.16)\quad Z(g(X,Y)) = g(\nabla_Z X, Y) + g(X, \nabla_Z Y) \qquad (\textit{metric compatible})$$

and

$$(2.17)\qquad [X,Y] = \nabla_X Y - \nabla_Y X \qquad (\textit{torsion free})$$

for all $X, Y, Z \in \mathfrak{X}(M)$. This connection ∇ is called the *Levi–Civita connection* of (M, g). As indicated above, (2.16) is the condition that the connection ∇ be compatible with the metric g, and (2.17) is the condition that ∇ be torsion free. Setting $Z = c'$ in (2.16), one finds that parallel translation of vector fields along any smooth curve c of M preserves inner products. For semi-Riemannian manifolds, the connection coefficients are given by

$$(2.18)\quad \Gamma_{ij}^k = \frac{1}{2}\sum_{a=1}^{n} g^{ak}\left(\frac{\partial g_{ia}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^a} + \frac{\partial g_{aj}}{\partial x^i}\right) \qquad (\textit{connection coefficients})$$

where g^{ij} represents the $(2,0)$ tensor defined by

$$(2.19)\qquad \sum_{a=1}^{n} g^{ia} g_{aj} = \delta^i{}_j \qquad \text{for } 1 \le i,j \le n.$$

The local representations g^{ij} and g_{ij} may be used to raise and lower indices. For example, if the upper index of the curvature tensor is lowered, one obtains the components of the *Riemann–Christoffel tensor* which is also known as the *covariant curvature tensor*.

$$(2.20)\qquad R_{ijkm} = \sum_{a=1}^{n} g_{ai} R^a{}_{jkm} \qquad (\textit{covariant curvature components})$$

Alternatively, one may define the Riemann–Christoffel tensor $\widetilde{R}$ as the $(0,4)$ tensor such that

$$(2.21)\qquad \widetilde{R}(W,Z,X,Y) = g(W, R(X,Y)Z).$$

Some standard curvature identities satisfied by the components of this tensor are $R_{ijkm} = R_{kmij} = -R_{jikm} = -R_{ijmk}$ and $R_{ijkm} + R_{ikmj} + R_{imjk} = 0$. The trace of the curvature tensor is the *Ricci curvature*, a symmetric $(0,2)$ tensor. For each $p \in M$, the Ricci curvature may be interpreted as a symmetric bilinear map $\text{Ric}_p : T_pM \times T_pM \to \mathbb{R}$. To evaluate $\text{Ric}(v,w)$, let $e_1, e_2, \ldots, e_n$ be an orthonormal basis for T_pM. Then

$$\text{Ric}(v,w) = \sum_{i=1}^{n} g(e_i, e_i)\, g(R(e_i, w)v, e_i) \tag{2.22}$$

or equivalently,

$$\text{Ric}(v,w) = \sum_{i=1}^{n} g(e_i, e_i)\, \widetilde{R}(e_i, v, e_i, w). \tag{2.23}$$

One may express v and w in the natural basis as $v = \sum v^i \partial/\partial x^i$ and $w = \sum w^i \partial/\partial x^i$ and then write

$$\text{Ric}(v,w) = \sum_{i,j=1}^{n} R_{ij}\, v^i w^j \tag{2.24}$$

where

$$R_{ij} = \sum_{a=1}^{n} R^a{}_{iaj}. \qquad (\textit{Ricci curvature components}) \tag{2.25}$$

If one uses the *Einstein summation convention* of summing over repeated indices, then equations (2.24) and (2.25) become $\text{Ric}(v,w) = R_{ij}v^i w^j$ and $R_{ij} = R^a{}_{iaj}$, respectively. The *Ricci tensor* is the $(1,1)$ tensor field which corresponds to the Ricci curvature. The components of the Ricci tensor may be obtained by raising one index of the Ricci curvature. Either index may be raised since the Ricci curvature is symmetric. Thus,

$$R^i{}_j = \sum_{a=1}^{n} g^{ai} R_{aj} = \sum_{a=1}^{n} g^{ai} R_{ja}. \qquad (\textit{Ricci tensor components}) \tag{2.26}$$

The trace of the Ricci curvature is the *scalar curvature* τ. Historically, this function has been denoted by the much used symbol R. Accordingly,

$$\tau = R = \sum_{a=1}^{n} R^a{}_a. \qquad (\textit{scalar curvature}) \tag{2.27}$$

Thus if $e_1, e_2, \ldots, e_n$ is an orthonormal basis of T_pM, one has

$$\tau = R = \sum_{i=1}^{n} g(e_i, e_i)\,\mathrm{Ric}(e_i, e_i). \tag{2.28}$$

The *gradient* and *Hessian* are defined for semi-Riemannian manifolds just as for Riemannian manifolds. If $f : M \to \mathbb{R}$ is a smooth function, then df is a $(0,1)$ tensor field (i.e., one-form) on M, and $\operatorname{grad} f$ is the $(1,0)$ tensor field (i.e., vector field) which corresponds to df. Thus,

$$Y(f) = df(Y) = g(\operatorname{grad} f, Y) \qquad (\textit{gradient}) \tag{2.29}$$

for an arbitrary vector field Y. In local coordinates $\big(U, (x^1, x^2, \ldots, x^n)\big)$, the vector field $\operatorname{grad} f$ is represented by

$$\operatorname{grad} f = \sum_{i,j=1}^{n} g^{ij} \frac{\partial f}{\partial x^i} \frac{\partial}{\partial x^j}. \qquad (\textit{gradient using coordinates}) \tag{2.30}$$

The *Hessian* H^f is defined to be the second covariant differential of f:

$$H^f = \nabla(\nabla f). \qquad (\textit{Hessian})$$

For a given $f \in \mathfrak{F}(M)$, the Hessian H^f is a symmetric (0,2) tensor field which is related to the gradient of f through the formula

$$H^f(X, Y) = g\big(\nabla_X(\operatorname{grad} f), Y\big)$$

for arbitrary vector fields X and Y. The *Laplacian* $\Delta f = \operatorname{div}(\operatorname{grad} f)$ is now defined to be the divergence of the gradient of f.

A tangent vector $v \in T_pM$ is classified as *timelike*, *nonspacelike*, *null*, or *spacelike* if $g(v, v)$ is negative, nonpositive, zero, or positive, respectively:

$$\begin{aligned} &g(v,v) < 0, && (\textit{timelike}) \\ &g(v,v) \le 0, && (\textit{nonspacelike} \text{ or } \textit{causal}) \\ &g(v,v) = 0, && (\textit{null} \text{ or } \textit{lightlike}) \\ &g(v,v) > 0. && (\textit{spacelike}) \end{aligned}$$

A *Riemannian manifold* (M, g) is a semi-Riemannian manifold of signature $(0, n)$ [i.e., $(+, \cdots, +)$]. Thus, on each tangent space T_pM of a Riemannian manifold the metric g_p is positive definite. Consequently, the metric induced on each tangent space is Euclidean, and all (nontrivial) vectors for Riemannian manifolds are spacelike.

A *Lorentzian manifold* is a semi-Riemannian manifold (M, g) of signature $(1, n-1)$ [i.e., $(-, +, \ldots, +)$]. At each point $p \in M$ the induced metric on the tangent space is Minkowskian. Each point of a Lorentzian manifold has timelike, null, and spacelike tangent vectors. A smooth curve is said to be timelike, null, or spacelike if its tangent vectors are always timelike, null, or spacelike, respectively.

A timelike curve in a Lorentzian manifold corresponds to the path of an observer moving at less than the speed of light. Null curves correspond to moving at the speed of light, and spacelike curves correspond to the geometric equivalent of moving faster than light. Although relativity predicts that physical particles cannot move faster than light, spacelike curves are of clear geometric interest.

A vector field X on M is timelike if $g(X, X) < 0$ at all points of M. A Lorentzian manifold with a given timelike vector field X is said to be *time oriented* by X. A *space–time* is a time oriented Lorentzian manifold. Not all Lorentzian manifolds may be time oriented, but a Lorentzian manifold which is not time orientable always admits a two-fold cover which is time orientable (cf. Chapter 3).

2.3 Null Cones and Semi-Riemannian Metrics

One of the folk theorems of general relativity asserts that the space–time metric is determined up to a conformal factor by the set of null vectors. In this section we examine more generally to what extent the null vectors of a nondefinite nondegenerate inner product on a vector space determine the given inner product and obtain as a consequence in Theorem 2.3 an elementary proof of this folk theorem for space–times [cf. Ehrlich (1991)].

Lemma 2.1. *Let V be a real n-dimensional vector space, $n \geq 2$, and let g and h be two nondefinite nondegenerate inner products on V of arbitrary signature. Suppose that g and h satisfy the condition that for any $v \in V$,*

$$g(v,v) = 0 \quad \text{iff} \quad h(v,v) = 0. \tag{2.31}$$

Then

(1) *either g and h or g and $-h$ have the same signature, and*
(2) *there exists $\lambda \neq 0$ such that*

$$h(v,w) = \lambda g(v,w) \tag{2.32}$$

for all $v, w \in V$.

Furthermore, if g has signature (s,r) with $r \neq s$, then $\lambda > 0$ if g and h have the same signature and $\lambda < 0$ if g and $-h$ have the same signature.

Proof. First consider the case that $r = s = 1$. Let $\{e_1, e_2\}$ be an arbitrary g-orthonormal basis for V such that $g(e_1, e_1) = -1$ and $g(e_2, e_2) = +1$. Then $\eta_1 = e_1 + e_2$ and $\eta_2 = e_1 - e_2$ are both g-null vectors. Hence, setting $h_{ij} = h(e_i, e_j)$, we obtain the system of equations

$$0 = h(\eta_1, \eta_1) = h_{11} + h_{22} + 2h_{12},$$
$$0 = h(\eta_2, \eta_2) = h_{11} + h_{22} - 2h_{12},$$

from which we conclude $h_{11} = -h_{22}$ and $h_{12} = 0$. Since $g(e_j, e_j) \neq 0$, we also have h_{11}, $h_{22} \neq 0$. Thus if we set $\lambda = -h_{11}$, then $h = \lambda g$ with $\lambda > 0$ if $h(e_1, e_1) < 0$ and $\lambda < 0$ if $h(e_1, e_1) > 0$.

Multiplying g by -1 if necessary, we consider the case that g has signature (r, s) with $r \geq 1$ and $s \geq 2$. Let $\{e_1, \ldots, e_r, e_{r+1}, \ldots, e_n\}$ be any g-orthonormal basis for V with $\{e_1, \ldots, e_r\}$ g-unit timelike and $\{e_{r+1}, \ldots, e_n\}$ g-unit spacelike. Again by (2.31), we have $h_{jj} = h(e_j, e_j) \neq 0$ for all j. Fix any $j, k \geq r+1$ with $j \neq k$. Then we have a one-parameter family of g-null vectors

$$v(\theta) = e_1 + \cos(\theta) \cdot e_j + \sin(\theta) \cdot e_k \tag{2.33}$$

for any $\theta \in \mathbb{R}$. In view of (2.31), we obtain

$$
\begin{aligned}
(2.34) \qquad 0 &= h(v(\theta), v(\theta)) \\
&= h_{11} + \cos^2(\theta) \cdot h_{jj} + \sin^2(\theta) \cdot h_{kk} \\
&\qquad + 2\cos(\theta) \cdot h_{1j} + 2\sin(\theta) \cdot h_{1k} + \sin(2\theta) \cdot h_{jk}.
\end{aligned}
$$

Taking $\theta = 0$ and $\theta = \pi$ respectively in (2.34) yields

$$
\begin{aligned}
&0 = h_{11} + h_{jj} + 2h_{1j}, \\
\text{and} \qquad &0 = h_{11} + h_{jj} - 2h_{1j},
\end{aligned}
$$

from which we conclude $h_{1j} = 0$ and $h_{jj} = -h_{11}$ for all $j \geq r + 1$. With this information, equation (2.34) reduces to the equation

$$0 = h_{11} - (\cos^2\theta + \sin^2\theta) \cdot h_{11} + \sin(2\theta) \cdot h_{jk}$$

or simply

$$0 = \sin(2\theta) \cdot h_{jk}.$$

Hence, $h_{jk} = 0$ for all $j, k \geq r + 1$, $j \neq k$.

We now have for $r = 1$ that $h_{pq} = 0$ if $p \neq q$, and $h_{11} = -h_{22} = -h_{33} = \cdots = -h_{nn}$. Hence if $\lambda = -h_{11} > 0$, then g and h both have signature $(-, +, \cdots, +)$, whereas if $\lambda = -h_{11} < 0$, then h has signature $(+, -, \cdots, -)$. In either case, $h = \lambda g$ with $\lambda = -h_{11}$ as required.

If $r \geq 2$, we have a little more work left. First, consideration of the one-parameter family

$$w(\theta) = \cos(\theta) \cdot e_j + \sin(\theta) \cdot e_k + e_n$$

with $j, k \leq r$ and $j \neq k$ yields $h_{jk} = h_{jn} = 0$ and $h_{11} = h_{22} = \cdots = h_{rr} = -h_{nn}$. It remains to show that $h_{1p} = 0$ and $h_{pq} = 0$ if $p \leq r < q$. But for this we need simply consider the one-parameter null family

$$x(\theta) = \cos(\theta) \cdot e_1 + \sin(\theta) \cdot e_p + e_q,$$

since the equation $0 = h(x(\theta), x(\theta))$ reduces to

$$0 = \sin(2\theta) \cdot h_{1p} + 2\sin(\theta) \cdot h_{pq}$$

in view of our above results. Thus we have obtained $h_{11} = h_{22} = \cdots = h_{rr} = -h_{r+1,r+1} = \cdots = -h_{nn}$ and $h_{jk} = 0$ if $j \neq k$, from which the desired conclusion follows as above. □

As a corollary, we obtain

Corollary 2.2. *Let V be a real n-dimensional vector space, $n \geq 3$, and assume V is equipped with inner products g and h, both of Lorentzian signature $(-,+,\cdots,+)$. Suppose g and h determine the same null vectors. Then $h = \lambda g$ for some constant $\lambda > 0$.*

With this last result applied to each tangent space, we obtain the desired result.

Theorem 2.3. *Let M be a smooth manifold of dimension $n \geq 3$ with metrics g and h, both of Lorentzian signature $(-,+,\cdots,+)$. Suppose for any $v \in TM$ that $g(v,v) = 0$ iff $h(v,v) = 0$. Then there exists a smooth function $\Omega : M \rightarrow (0,\infty)$ such that $h = \Omega g$.*

A somewhat more general class of objects than globally conformally related metrics has been considered in differential geometry and general relativity, namely, conformal transformations. Here a global diffeomorphism $F : (M,g) \rightarrow (N,h)$ between two semi-Riemannian manifolds (M,g) and (N,h) is said to be a *global conformal transformation* if there exists a smooth function $\Omega : M \rightarrow (0,\infty)$ such that

$$h(F_* v, F_* w) = \Omega(p)\, g(v,w) \tag{2.35}$$

for all $v, w \in T_pM$ and $p \in M$. In the Riemannian case, such maps would be angle preserving. Evidently, condition (2.35) implies that (M,g) and (N,h) have the same signature.

Conversely, given two semi-Riemannian manifolds (M,g) and (N,h), we may apply Lemma 2.1 to g and F^*h in order to obtain that the condition

$$g(v,v) = 0 \quad \text{iff} \quad h(F_* v, F_* v) = 0 \qquad \text{for all } v \in TM \tag{2.36}$$

serves to ensure that F is a global conformal transformation of (M,g) onto either (N,h) or $(N,-h)$. Especially for space–times, one has the following well-known extension of Theorem 2.3.

Corollary 2.4. *Let (M, g) and (N, h) be smooth manifolds of dimension $n \geq 3$ having Lorentzian signature $(-, +, \cdots, +)$. Suppose that $f : M \to N$ is a diffeomorphism which satisfies condition* (2.36). *Then f is a global conformal transformation of (M, g) onto (N, h).*

One could paraphrase this last result as follows: a diffeomorphism of space–times which preserves null vectors is a conformal transformation. Hence Corollary 2.4 is a differentiable version of a much deeper a priori topological result of Hawking, King, and McCarthy (1976).

2.4 Sectional Curvature

Let (M, g) be a semi-Riemannian manifold. A two-dimensional linear subspace E of T_pM is called a *plane section*. The plane section E is said to be *nondegenerate* if for each nontrivial vector $v \in E$ there is some vector $u \in E$ with $g(u, v) \neq 0$. This is equivalent to the requirement that $g_p \,|\, E$ be a nondegenerate inner product on E. If v and w form a basis of the plane section E, then $g(v, v)g(w, w) - [g(v, w)]^2$ is a nonzero quantity if and only if E is nondegenerate. This quantity represents the square of the semi-Euclidean area of the parallelogram determined by v and w. The plane E is said to be *timelike* if the signature of $g_p \,|\, E$ is $(1, 1)$, i.e., $(-, +)$. It is *spacelike* if the signature is $(0, 2)$, i.e., $(+, +)$.

For Lorentzian manifolds, degenerate planes are called either *null* or *lightlike* and always have signature $(0, +)$. A null plane in T_pM is a plane tangent to the null cone in T_pM. Thus, in the Lorentzian case a degenerate plane contains exactly one generator of the null cone.

Using the square of the semi-Euclidean area of the parallelogram determined by the basis vectors $\{v, w\}$, one has the following classification for plane sections of Lorentzian manifolds:

$$g(v, v)g(w, w) - [g(v, w)]^2 < 0, \qquad (\textit{timelike plane})$$

$$g(v, v)g(w, w) - [g(v, w)]^2 = 0, \qquad (\textit{degenerate plane})$$

$$g(v, v)g(w, w) - [g(v, w)]^2 > 0. \qquad (\textit{spacelike plane})$$

The *sectional curvature* of the nondegenerate plane section E with basis

$\{v, w\}$ where $v = \sum v^i \partial/\partial x^i$ and $w = \sum w^i \partial/\partial x^i$ is then given by

$$
\begin{aligned}
K(p, E) &= \frac{g(R(w,v)v,w)}{g(v,v)g(w,w) - [g(v,w)]^2} \\
&= \frac{\widetilde{R}(w,v,w,v)}{g(v,v)g(w,w) - [g(v,w)]^2} \qquad (\textit{sectional curvature}) \\
&= \frac{\sum R_{ijkm} w^i v^j w^k v^m}{\sum g_{ij} v^i v^j g_{km} w^k w^m - [\sum g_{ij} v^i w^j]^2}.
\end{aligned} \tag{2.37}
$$

For positive definite manifolds the Ricci curvature evaluated at a unit vector is sometimes thought of as more or less an average sectional curvature weighted by a factor of $(n-1)$. More precisely, let w be a unit vector at p in the *Riemannian* manifold (M, g), extend to an orthonormal basis $\{e_1, e_2, \ldots, e_{n-1}, e_n = w\}$, and let $E_i =$ span$\{e_i, w\}$ for $1 \leq i \leq n-1$. Equations (2.22) and (2.37) yield

$$\operatorname{Ric}(w,w) = \sum_{i=1}^{n-1} g(R(e_i,w)w,e_i) = \sum_{i=1}^{n-1} K(p,E_i).$$

It is instructive to contrast the above with the Ricci curvature evaluated at a unit timelike vector in a Lorentzian manifold. In this case the interpretation is in terms of the negative of the timelike sectional curvature. For let u be a unit *timelike* vector at a point p of the *Lorentzian* manifold (M, g). Extend to an orthonormal basis $\{e_1, e_2, \ldots, e_{n-1}, e_n = u\}$ and let $E_i =$ span$\{e_i, u\}$ for $1 \leq i \leq n-1$. Equations (2.22) and (2.37) yield

$$\operatorname{Ric}(u,u) = \sum_{i=1}^{n-1} g(R(e_i,u)u,e_i) = -\sum_{i=1}^{n-1} K(p,E_i).$$

Thus, if u is a timelike vector with $\operatorname{Ric}(u,u) > 0$, then in some sense the "average" sectional curvature for planes in the pencil of u is negative.

For Riemannian manifolds one has a number important "pinching" theorems [cf. Cheeger and Ebin (1975)]. However, similar results fail for Lorentzian manifolds. In particular, if the sectional curvatures of timelike planes are bounded *both* above and below, and if $\dim(M) \geq 3$, then (M, g) has constant curvature [cf. Harris (1982a), Dajczer and Nomizu (1980a)]. Nevertheless, families of Lorentzian manifolds conformal to ones of constant curvature may be constructed which have all timelike sectional curvatures bounded in one direction [cf. Harris (1979)]. However, if $\dim(M) \geq 3$ and if the sectional curvatures of *all* nondegenerate planes are bounded either from above or from below, then the sectional curvature is constant [cf. Kulkarni (1979)]. Sectional curvature has been further investigated by Dajczer and Nomizu (1980b), Nomizu (1983), Beem and Parker (1984), and Cordero and Parker (1995a,c).

A semi-Riemannian manifold (M, g) which has the same sectional curvature on all (nondegenerate) sections is said to have *constant curvature*. If (M, g) has constant curvature c, then $R(X,Y)Z = c\,[\,g(Y,Z)X - g(X,Z)Y]$ [cf. Graves and Nomizu (1978)]. Thorpe (1969) showed that the sectional curvature can only be continuously extended to degenerate planes in the case of constant curvature. Spaces of constant sectional curvature have been investigated in connection with the space form problem [cf. Wolf (1961, 1974)].

The curvature tensor, Ricci curvature, scalar curvature, and sectional curvature may all be calculated in local coordinates using the metric tensor components and the first two partial derivatives of these components. Thus, the metric tensor determines the curvatures. In contrast, curvature does not necessarily determine the metric. Nevertheless, for most Lorentzian manifolds the metric will be determined either completely or up to a constant conformal factor given sufficient information concerning curvature in local coordinates [cf. Hall (1983, 1984), Hall and Kay (1988), Ihrig (1975), Quevedo (1992)].

Conjugate points along geodesics may be defined using Jacobi fields. These objects are studied in Chapter 10 in connection with the development of the Morse index theory for timelike and null geodesics. If $c : (a, b) \to M$ is a geodesic, then the smooth vector field $J : (a, b) \to TM$ along c is a *Jacobi field* if it satisfies the *Jacobi equation*,

$$J'' + R(J, c')c' = 0, \qquad \text{(Jacobi equation)} \tag{2.38}$$

where $J'' = \nabla_{c'}(\nabla_{c'} J)$. In an intuitive sense, one thinks of a Jacobi field as representing the relative displacement of "nearby" geodesics [cf. Hawking and Ellis (1973), Hicks (1965), or Misner, Thorne and Wheeler (1973)]. In particular, let c be a unit speed timelike geodesic. Then c represents the path of a "freely falling" particle moving at less than the speed of light. Taking J

as a Jacobi field which is orthogonal to the tangent vector c', one interprets J as a vector from the original particle to another particle moving on a nearby timelike geodesic, and one interprets J'' as the relative (or tidal) acceleration of the second particle as measured by the first. If $g(J,J) \neq 0$, then the definition of sectional curvature together with $g(c',c') = -1$ and $g(J,c') = 0$ yield

$$K(c',J) = \frac{g(R(J,c')c',J)}{-g(J,J)} = \frac{g(-R(J,c')c',J)}{g(J,J)}.$$

At points where J does not vanish, the vector $J/\sqrt{g(J,J)}$ is a unit vector in the direction of J. Using $J'' = -R(J,c')c'$, one finds that the *radial component* of the tidal acceleration is given by

$$\begin{aligned} \frac{g(J'',J)}{\sqrt{g(J,J)}} &= \frac{g(-R(J,c')c',J)}{\sqrt{g(J,J)}} \\ &= \left(\frac{g(-R(J,c')c',J)}{g(J,J)}\right)\sqrt{g(J,J)} \\ &= K(c',J)\sqrt{g(J,J)} \\ &= K(c',J)\,|J|. \end{aligned} \tag{2.39}$$

This equation shows that for "close" particles the radial component of the tidal acceleration varies directly with the separation distance $|J|$ *and with the sectional curvature* $K(c',J)$ *of the plane containing both* c' *and* J. Thus, *using our sign conventions*, a timelike plane with positive sectional curvature corresponds to freely falling particles accelerating away from each other, and negative sectional curvature of a timelike plane corresponds to particles accelerating toward each other. Since $\text{Ric}(c',c') > 0$ corresponds to the timelike planes containing c' having negative average sectional curvature, it follows that $\text{Ric}(c',c') > 0$ *corresponds to average attractive* (i.e., *focusing*) *tidal forces*. It should be kept in mind that some authors use different sign conventions and may have sectional curvature equal to the negative of ours.

For a constant value of $|J|$ the maximum tidal acceleration will be radial [cf. Beem and Parker (1990, p. 820)]. Thus *at any fixed value* t_0, *an observer traversing the timelike geodesic* c *will have zero tidal accelerations if and only if all planes* E *containing* $c'(t_0)$ *have zero sectional curvature.*

2.5 The Generic Condition

The sectional curvature can be used to study the generic condition, which will be of importance in the singularity theorems to be considered in Chapter 12. If $W = \sum_{a=1}^{n} W^a \, \partial/\partial x^a$ is a tangent vector, then the values W^a are the *contravariant* components. Using the metric g one has values $W_b = \sum_{a=1}^{n} g_{ab} W^a$ which are the *covariant* components. Thus, $\sum_{b=1}^{n} W_b \, dx^b$ is the cotangent vector corresponding to the original W. The *generic condition* is said to be satisfied for a vector W at $p \in M$ if

$$\sum_{c,d=1}^{n} W^c W^d W_{[a} R_{b]cd[e} W_{f]} \neq 0. \qquad (\textit{generic condition}) \tag{2.40}$$

If condition (2.40) fails to hold (i.e., if $\sum_{c,d=1}^{n} W^c W^d W_{[a} R_{b]cd[e} W_{f]} = 0$), then W will be called *nongeneric*.

The *generic condition is said to hold for a geodesic* $c : (a, b) \to M$ *if at some point* $c(t_0)$ *the tangent vector to the geodesic is generic, i.e., one has* (2.40) *satisfied with* $W = c'(t_0)$. Notice from continuity that if (2.40) holds for some $W = c'(t_0)$, then it will hold for $W = c'(t)$ whenever t is sufficiently close to t_0.

We will show in Proposition 2.6 that for a vector W which is nonnull (i.e., $g(W, W) \neq 0$), the nongeneric condition is equivalent to requiring that all plane sections containing W have zero sectional curvature. For null vectors, the nongeneric condition is slightly more complicated.

Lemma 2.5. *Let* R_{abcd} *represent the components of the curvature tensor with respect to an orthonormal basis* $\{v_1, v_2, \ldots, v_n\}$ *of* T_pM. *The vector* $W = v_n$ *satisfies the generic condition* (2.40) *iff there exist* b *and* e *with* $1 \leq b, e \leq n-1$ *such that* $R_{bnne} \neq 0$.

Proof. The components of W are given by $W^1 = \cdots = W^{n-1} = W_1 = \cdots = W_{n-1} = 0$, $W^n = 1$, and $W_n = +1$ or -1. Consequently,

$$\sum_{c,d=1}^{n} W^c W^d W_{[a} R_{b]cd[e} W_{f]} = \frac{1}{4}\big(W_a\, R_{bnne}\, W_f - W_b\, R_{anne}\, W_f$$
$$- W_a\, R_{bnnf}\, W_e + W_b\, R_{annf}\, W_e\big)$$
$$= \frac{1}{4}\big(\delta^n{}_a\, R_{bnne}\, \delta^n{}_f - \delta^n{}_b\, R_{anne}\, \delta^n{}_f \tag{2.41}$$
$$- \delta^n{}_a\, R_{bnnf}\, \delta^n{}_e + \delta^n{}_b\, R_{annf}\, \delta^n{}_e\big).$$

It is easily seen that this expression is nonzero if and only if $R_{bnne} \neq 0$ for some b, e with $1 \leq b, e \leq n-1$. □

Proposition 2.6. *If $W \in T_pM$ is a nonnull vector, then W fails to be generic (i.e., is nongeneric) iff for each nondegenerate plane section E containing W the sectional curvature $K(p, E)$ vanishes.*

Proof. We may assume without loss of generality that W is a unit vector. Set $W = v_n$ and extend to an orthonormal basis $\{v_1, v_2, \ldots, v_n\}$ of T_pM. Let R_{abcd} represent the components of the curvature tensor with respect to this basis.

From Lemma 2.5, if W fails to satisfy the generic condition, then $R_{bnne} = 0$ for all $1 \leq b, e \leq n-1$. This and the skew symmetry of R_{abcd} in both the first pair of indices and the second pair of indices yield $R_{bnne} = 0$ for all $1 \leq b, e \leq n$. Hence $R_{nbne} = -R_{bnne} = 0$ for all such b and e. It follows that if U is another tangent vector at p, one has $\sum_{a,b,c,e=1}^{n} R_{abce} W^a U^b W^c U^e = \sum_{b,e=1}^{n} R_{nbne} U^b U^e = 0$. From (2.37), we conclude that if E is the plane spanned by $\{W, U\}$, and if E is nondegenerate, then $K(p, E) = 0$.

Conversely, assume all nondegenerate planes containing W have zero sectional curvature. Then (2.37) easily shows that all terms of the form R_{bnnb} must be zero. Assume $b \neq e$. Notice that $E = \operatorname{span}\{v_b, v_b + 2v_e\}$ cannot be degenerate. Using (2.37) we obtain $\widetilde{R}(v_n, v_b, v_n, v_b) = \widetilde{R}(v_n, v_e, v_n, v_e) = \widetilde{R}(v_n, v_b + 2v_e, v_n, v_b + 2v_e) = 0$. The multilinearity of $\widetilde{R}$ and standard curvature identities then yield $0 = \widetilde{R}(v_n, v_b + 2v_e, v_n, v_b + 2v_e) = 4\,\widetilde{R}(v_n, v_b, v_n, v_e)$. This shows $R_{nbne} = 0$ and hence $R_{bnne} = 0$. Using Lemma 2.5, it follows that W fails to satisfy the generic condition as desired. □

Proposition 2.6 implies that the only way for a timelike geodesic c to fail to satisfy the generic condition is for the corresponding observer to fail to *ever* experience any tidal accelerations.

In Chapter 12 we will make use of an alternative formulation of the generic condition using the curvature tensor. Let c be a unit speed timelike geodesic, $p = c(t_0)$, and $W = c'(t_0)$. The set of vectors orthogonal to W is an $(n-1)$-dimensional linear subspace $W^\perp = V^\perp(c(t_0))$ lying in T_pM, and the metric induced on this linear subspace is positive definite. Thus $V^\perp(c(t_0))$ is a spacelike hyperplane in T_pM. If $y \in T_pM$, then $g\big(W, R(y,W)W\big) = 0$ which implies that $R(y,W)W$ lies in $V^\perp(c(t_0))$. It follows that the curvature tensor R induces a linear map from $V^\perp(c(t_0))$ to $V^\perp(c(t_0))$:

$$R(\,\cdot\,,W)W|_{t_0} : V^\perp(c(t_0)) \to V^\perp(c(t_0)). \tag{2.42}$$

Since $g \mid V^\perp(c(t_0))$ is positive definite, this curvature map is nontrivial if and only if there are vectors $y_1, y_2 \in V^\perp(c(t_0))$ with $g\big(y_2, R(y_1,W)W\big) \neq 0$. The next result shows this map is nontrivial if and only if W is generic.

Proposition 2.7. *If $W = c'(t_0)$ is a timelike vector in the Lorentzian manifold (M,g), then the following three conditions are equivalent:*

(1) *The timelike vector W is generic.*

(2) *At least one plane containing W has nonzero sectional curvature.*

(3) *$R(\,\cdot\,,W)W$ is not the trivial map.*

Proof. Clearly, Proposition 2.6 shows the first two conditions are equivalent. Let R_{abcd} represent the components of the covariant curvature tensor with respect to an orthonormal basis $\{v_1, v_2, \ldots v_{n-1}, v_n = W\}$ of T_pM. If W satisfies the generic condition, then Lemma 2.5 implies that $-R_{bnen} = R_{bnne} \neq 0$ for some $1 \leq b, e \leq n-1$. Consequently, $g(v_b, R(v_e, v_n)v_n) \neq 0$ which shows $R(v_e, v_n)v_n$ is a nontrivial vector. Thus, when W is generic the map $R(\,\cdot\,,W)W$ is not the trivial map.

Conversely, assume that $R(\,\cdot\,,W)W$ is not trivial. Let y_1, y_2 be vectors in $W^\perp = \operatorname{span}\{v_1, v_2, \ldots, v_{n-1}\}$ with $g(y_2, R(y_1, v_n)v_n) \neq 0$. This last inequality together with the multilinearity of $\widetilde{R}(W,Z,X,Y) = g(W, R(X,Y)Z)$ yield b

and e with $1 \leq b, e \leq n-1$ such that $R_{bnne} \neq 0$. Thus, Lemma 2.5 shows that W is generic as desired. $\square$

The next proposition shows that a sufficient condition for the nonnull vector W to be generic is that $\mathrm{Ric}(W, W) \neq 0$.

Proposition 2.8. *If $W \in T_pM$ is a nonnull vector with* $\mathrm{Ric}(W, W) \neq 0$, *then W is generic.*

Proof. We may assume without loss of generality that W is a unit vector and extend to an orthonormal basis $\{v_1, v_2, \ldots, v_{n-1}, v_n = W\}$ of T_pM. Then,

$$\begin{aligned} \mathrm{Ric}(W, W) &= \mathrm{Ric}(v_n, v_n) \\ &= \sum_{i=1}^{n} g(v_i, v_i)\, g(R(v_i, v_n)v_n, v_i) \\ &= \sum_{i=1}^{n-1} g(v_i, v_i)\, R_{inin} \end{aligned}$$

which shows that $\mathrm{Ric}(W, W) \neq 0$ implies $R_{inin} \neq 0$ for some $1 \leq i \leq n-1$. The result now follows from Lemma 2.5. $\square$

In particular, this last proposition shows that if $c : (a, b) \to M$ is a timelike geodesic with $\mathrm{Ric}(c', c') > 0$ for some $t = t_0$, then c is generic.

In order to investigate the generic condition for (nontrivial) null vectors, we first define a *pseudo-orthonormal basis* [cf. Hawking and Ellis (1973)]. Let $\{v_1, v_2, \ldots, v_{n-2}\}$ be $n-2$ unit spacelike vectors in T_pM that are mutually orthogonal, and let W, N be two null vectors which satisfy $g(W, N) = -1$ and which are both orthogonal to the first $n-2$ vectors. Then a pseudo-orthonormal basis of T_pM is formed by $\{v_1, v_2, \ldots, v_{n-2}, N, W\}$. In this basis the metric tensor at p has the form

$$g_{ij} = \delta_{ij} \qquad \text{and} \qquad g_{i\,n-1} = g_{in} = 0 \qquad \text{for } 1 \leq i, j \leq n-2,$$

$$\text{with} \qquad g_{n\,n-1} = -1 \qquad \text{and} \qquad g_{nn} = g_{n-1\,n-1} = 0.$$

It is not difficult to show that any nontrivial null vector W can be extended to a pseudo-orthonormal basis. One first takes any timelike plane E containing

W and finds an orthonormal basis $\{e_{n-1}, e_n\}$ of E with $W = (e_{n-1} + e_n)/\sqrt{2}$, e_{n-1} spacelike, and e_n timelike. Then one extends to an orthonormal basis $\{e_1, e_2, \ldots, e_n\}$ and sets $v_i = e_i$ for $1 \leq i \leq n-2$. Finally, one assigns $v_{n-1} = N = (-e_{n-1} + e_n)/\sqrt{2}$, and $v_n = W$.

The next lemma provides the null version of Lemma 2.5 using a pseudo-orthonormal basis. In the lemma below, the values of b and e only go to $n-2$ as opposed to $n-1$ in Lemma 2.5.

Lemma 2.9. *Let R_{abcd} represent the components of the curvature tensor with respect to a pseudo-orthonormal basis $\{v_1, \ldots, v_{n-2}, v_{n-1} = N, v_n = W\}$ of T_pM. The null vector $W = v_n$ satisfies the generic condition* (2.40) *iff there exist b and e with $1 \leq b, e \leq n-2$ such that $R_{bnne} \neq 0$.*

Proof. The components of W are given by $W^1 = \cdots = W^{n-1} = W_1 = \cdots = W_{n-2} = W_n = 0$, $W^n = 1$, and $W_{n-1} = -1$. Consequently,

$$\begin{aligned}\sum_{c,d=1}^{n} W^c W^d W_{[a} R_{b]cd[e} W_{f]} &= \frac{1}{4}\big(W_a\, R_{bnne}\, W_f - W_b\, R_{anne}\, W_f \\ &\qquad - W_a\, R_{bnnf}\, W_e + W_b\, R_{annf}\, W_e\big) \\ &= \frac{1}{4}\big(\delta^{n-1}{}_a\, R_{bnne}\, \delta^{n-1}{}_f - \delta^{n-1}{}_b\, R_{anne}\, \delta^{n-1}{}_f \\ &\qquad - \delta^{n-1}{}_a\, R_{bnnf}\, \delta^{n-1}{}_e + \delta^{n-1}{}_b\, R_{annf}\, \delta^{n-1}{}_e\big).\end{aligned}$$

It is easily seen that this expression is zero for all a, b, e, f whenever one or more of a, b, e, f equal n. If it is nonzero for some a, b, e, f, then exactly one of a, b must be $n-1$, and also exactly one of e, f must be $n-1$. It follows that $\sum_{c,d=1}^{n} W^c W^d W_{[a} R_{b]cd[e} W_{f]} \neq 0$ iff $R_{bnne} \neq 0$ for some $1 \leq b, e \leq n-2$. □

If $n = 2$, then Lemma 2.9 shows all null vectors fail to be generic since there are no values of b and e with $1 \leq b, e \leq n-2$.

Corollary 2.10. *If* $\dim(M) = 2$, *then all null vectors are nongeneric.*

If W is a nontrivial null vector in T_pM, then the orthogonal space $W^\perp = N(W) = \{v \in T_pM \mid g(W, v) = 0\}$ is an $(n-1)$-dimensional linear subspace of the tangent space and contains the null vector W. The set $N(W)$ is a hyperplane which is tangent to the null cone at p along one generator. The signature of $g \mid N(W)$ is degenerate of order one and positive of order $n-2$. Notice that $g(W, R(v,W)W) = 0$ implies that $R(v,W)W$ lies in $N(W)$.

The one-dimensional linear subspace determined by W will be denoted by $[W]$. Let $G(W) = N(W)/[W]$ be the $(n-2)$-dimensional quotient space and $\pi : N(W) \to G(W)$ be the natural projection map. Since $R(W,W) = 0$ and the multilinearity of the curvature tensor yield $R(v+aW, W)W = R(v,W)W$ for all v, each element of a given coset of $G(W)$ is mapped to the same element of $N(W)$ by $R(\,\cdot\,, W)W$. It follows that one has a linear map

$$\overline{R}(\,\cdot\,, W)W : G(W) \to G(W) \tag{2.43}$$

defined by $\overline{R}\,(\overline{v}, W)W = \pi\,(R(v,W)W)$ where v is any vector in $\pi^{-1}(\overline{v})$. The degenerate metric $g \mid N(W)$ projects to a positive definite metric $\overline{g}$ on $G(W)$ since $v_1, v_2 \in N(W)$ must satisfy $g(v_1 + aW, v_2 + bW) = g(v_1, v_2)$ for all real a, b. Thus,

$$\overline{g}(\overline{v}_1, \overline{v}_2) = g(v_1, v_2) \qquad (\textit{quotient metric on } G(W)) \tag{2.44}$$

where $\overline{v}_1 = \pi(v_1)$ and $\overline{v}_2 = \pi(v_2)$.

The next result is the null version of Proposition 2.7. We do not give a statement corresponding to condition (2) of Proposition 2.7. In general, a null vector may lie in planes with nonzero sectional curvature and yet fail to be generic.

Proposition 2.11. *Let (M,g) be a Lorentzian manifold, and let $W \in T_pM$ be a nontrivial null vector. Then the following two conditions are equivalent:*

(1) *The null vector W is generic.*
(2) *$\overline{R}(\,\cdot\,, W)W$ is not the trivial map.*

Proof. Let R_{abcd} represent the components of the covariant curvature tensor with respect to a pseudo-orthonormal basis $\{v_1, \ldots, v_{n-2}, v_{n-1}, v_n = W\}$ of T_pM, $N(W) = W^\perp$, and $G(W) = N(W)/[W]$ as above. Notice that $N(W) = \operatorname{span}\{v_1, v_2, \ldots, v_{n-2}, W\}$.

Assuming W is generic, Lemma 2.9 yields b and e with $R_{bnne} \neq 0$ for $1 \leq b, e \leq n-2$. Consequently, $\overline{g}\,\left(\pi(v_b)\,, \overline{R}(\pi(v_e), W)W\right) = g(v_b, R(v_e, W)W) = R_{bnen} = -R_{bnne} \neq 0$ which shows that $\overline{R}(\,\cdot\,, W)W$ is not trivial.

Conversely, assume $\overline{R}(\cdot, W)W$ is not the trivial map. Since $\overline{g}$ is positive definite, there must exist $\overline{u}$ and $\overline{v}$ in $G(W)$ with $\overline{g}(\overline{u}, \overline{R}(\overline{v}, W)W) \neq 0$. Choose $u \in \pi^{-1}(\overline{u})$ and $v \in \pi^{-1}(\overline{v})$. Then multilinearity, the fact that $\overline{g}(\overline{u}, \overline{R}(\overline{v}, W)W) = g(u, R(v, W)W) \neq 0$, and $R(W, W)W = 0$ together imply there must be b and e, $1 \leq b, e \leq n-2$, such that $g(v_b, R(v_e, W)W) \neq 0$. Hence, $R_{bnen} = -R_{bnne} \neq 0$, and W is generic by Lemma 2.9. □

The next proposition shows that $\text{Ric}(W, W) \neq 0$ is a sufficient condition for a vector W to be generic. This was already proven for timelike and spacelike vectors in Proposition 2.8.

Proposition 2.12. *Let (M, g) be a Lorentzian manifold, and let $W \in T_pM$ be a tangent vector. If* $\text{Ric}(W, W) \neq 0$, *then W is generic.*

Proof. We may assume without loss of generality that W is a null vector because of Proposition 2.8. It is always possible to construct both an orthonormal basis $\{e_1, e_2, \ldots, e_n\}$ and a pseudo-orthonormal basis

$$\{v_1, \ldots, v_{n-2}, v_{n-1} = N, v_n = W\}$$

with $v_i = e_i$ for $1 \leq i \leq n-2$,

$$N = (-e_{n-1} + e_n)/\sqrt{2}, \qquad W = (e_{n-1} + e_n)/\sqrt{2},$$

and e_n timelike. Let R_{abcd} be the components of the Riemann–Christoffel curvature tensor with respect to the pseudo-orthonormal basis and $\overline{R}_{abcd}$ be the components with respect to the orthonormal basis. Then,

$$\begin{aligned}
\sum_{i=n-1}^{n} g(e_i, e_i)\, g(R(e_i, W)W, e_i) &= \frac{1}{2}[g(R(e_{n-1}, e_{n-1} + e_n)(e_{n-1} + e_n), e_{n-1}) \\
&\qquad - g(R(e_n, e_{n-1} + e_n)(e_{n-1} + e_n), e_n)] \\
&= \frac{1}{2}[g(R(e_{n-1}, e_n)e_n, e_{n-1}) \\
&\qquad - g(R(e_n, e_{n-1})e_{n-1}, e_n)] \\
&= \frac{1}{2}\left(\overline{R}_{n-1\,n\,n-1\,n} - \overline{R}_{n\,n-1\,n\,n-1}\right) = 0.
\end{aligned}$$

Hence,

$$\begin{aligned}\operatorname{Ric}(W,W) &= \sum_{i=1}^{n} g(e_i,e_i)\, g(R(e_i,W)W,e_i)\\ &= \sum_{i=1}^{n-2} g(e_i,e_i)\, g(R(e_i,W)W,e_i)\\ &= \sum_{i=1}^{n-2} g(v_i,v_i)\, g(R(v_i,v_n)v_n,v_i)\\ &= \sum_{i=1}^{n-2} g(v_i,v_i)\, R_{inin},\end{aligned}$$

which shows that $\operatorname{Ric}(W,W) \neq 0$ implies $R_{inin} \neq 0$ for some $1 \leq i \leq n-2$. The result now follows from Lemma 2.9. □

The next corollary follows from Corollary 2.10 and Proposition 2.12.

Corollary 2.13. *Let (M,g) be a Lorentzian manifold with* $\dim(M) = 2$. *If W is null, then* $\operatorname{Ric}(W,W) = 0$.

The generic condition holds generically in each tangent space of any point where there is some component of the curvature tensor not equal to zero. More precisely, one can establish the following result [cf. Beem and Harris (1993a, p. 950)].

Proposition 2.14. *Let (M,g) be a Lorentzian manifold of dimension four and let $p \in M$. If T_pM has five nonnull and nongeneric vectors with four of them linearly independent and with the fifth not in any plane determined by any two of the original four, then the curvature tensor vanishes at p. In particular, if any component of the curvature tensor fails to be zero at p, then one cannot find five nonnull nongeneric vectors at p in general position.*

The situation for nongeneric null vectors is somewhat different. Having all null vectors nongeneric does not necessarily imply zero curvature at a point. For dimension three and higher, all null vectors at a point will be nongeneric if and only if there is constant sectional curvature at the point. In dimension four, one can have a "cubic" of nongeneric null vectors at a point and yet fail to have constant sectional curvature at that point. Using a cubic definition for "generically situated" [cf. Beem and Harris (1993b, pp. 969–972)], one may establish the following result.

Proposition 2.15. *Let (M, g) be a Lorentzian manifold of dimension four, and let $p \in M$. If T_pM has 11 null nongeneric vectors generically situated, then all sectional curvatures at p are equal. In particular, if (M, g) does not have constant sectional curvature at p, then the generic null directions at p form an open dense subset of the two-sphere of all null directions at p.*

This last proposition yields that "generically" a given null direction at a point p satisfies the generic condition.

Assume that (M, g) is a four-dimensional Lorentzian manifold, and recall that the Jacobi equation is $J'' + R(J, c')c' = 0$. Let c be a unit speed timelike vector, and assume $\{E_1, E_2, E_3, E_4\}$ is an orthonormal basis along c which moves by parallel translation along c and satisfies $E_4 = c'(t)$. If $J = \sum_{i=1}^{4} J^i E_i$ is a Jacobi field, the Jacobi equation can then be written as

$$\begin{aligned}\frac{d^2J^i}{dt^2} &= -\sum_{k=1}^{4} R^i{}_{4k4} J^k \\ &= -\sum_{k=1}^{3} R^i{}_{4k4} J^k \qquad (\textit{Jacobi equation})\end{aligned}$$

using $R^i{}_{444} = 0$. Since $\{E_1, E_2, E_3, E_4\}$ is an orthonormal basis with $c'(t) = E_4$, the vectors $\{E_1, E_2, E_3\}$ are unit spacelike vectors, and the Jacobi equation becomes

$$\frac{d^2J^i}{dt^2} = -\sum_{k=1}^{3} R_{i4k4} J^k. \qquad (\textit{using an orthonormal basis})$$

We now illustrate the unboundedness of tidal accelerations for observers with velocity vectors close to the direction of a null vector W which satisfies the generic condition [cf. Beem and Parker (1990)]. To this end, assume that (U, x_1, x_2, x_3, x_4) are local coordinates near $c(t_0) = p$. Assume also that the natural basis for the x_i-coordinates at p is the orthonormal set $\{E_1, E_2, E_3, E_4\}$. Denote the components of the curvature tensor for this basis by R_{abcd}. Regard θ as temporarily fixed, and define new coordinates (y_1, y_2, y_3, y_4) near p by $y_1 = x_1$, $y_2 = x_2$, $y_3 = x_3 \cosh(\theta) - x_4 \sinh(\theta)$, and $y_4 = -x_3 \sinh(\theta) + x_4 \cosh(\theta)$. Let the natural basis at p for the y-coordinates be denoted by $\{\overline{E}_1, \overline{E}_2, \overline{E}_3, \overline{E}_4\}$, and let the curvature tensor components at p with respect to this basis be denoted by $\overline{R}_{abcd}$. The new basis $\{\overline{E}_1, \overline{E}_2, \overline{E}_3, \overline{E}_4\}$ is also orthonormal, and if one lets $\theta \to +\infty$, the direction of $\overline{E}_4$ converges to the null direction determined by $W = E_3 + E_4$. This change of coordinates corresponds in T_pM to what is called a "pure boost." The new curvature components $\overline{R}_{abcd}$ are related to the original components R_{abcd} by

$$\overline{R}_{ijkm} = \sum_{a,b,c,d=1}^{4} R_{abcd} \frac{\partial x^a}{\partial y^i} \frac{\partial x^b}{\partial y^j} \frac{\partial x^c}{\partial y^k} \frac{\partial x^d}{\partial y^m}.$$

Using $x_1 = y_1$, $x_2 = y_2$, $x_3 = y_3 \cosh(\theta) + y_4 \sinh(\theta)$, and $x_4 = y_3 \sinh(\theta) + y_4 \cosh(\theta)$, one finds

$$\overline{R}_{3434} = R_{3434}, \tag{2.45}$$

$$\overline{R}_{3424} = \cosh(\theta) R_{3424} + \sinh(\theta) R_{3423}, \tag{2.46}$$

$$\overline{R}_{3414} = \cosh(\theta) R_{3414} + \sinh(\theta) R_{3413}, \tag{2.47}$$

$$\begin{aligned}\overline{R}_{2424} &= \cosh^2(\theta) R_{2424} + 2\cosh(\theta)\sinh(\theta) R_{2423} \\ &\quad + \sinh^2(\theta) R_{2323},\end{aligned} \tag{2.48}$$

$$\begin{aligned}\overline{R}_{2414} &= \cosh^2(\theta) R_{2414} + \cosh(\theta)\sinh(\theta) R_{2413} \\ &\quad + \cosh(\theta)\sinh(\theta) R_{2314} + \sinh^2(\theta) R_{2313},\end{aligned} \tag{2.49}$$

$$\begin{aligned}\overline{R}_{1414} &= \cosh^2(\theta) R_{1414} + 2\cosh(\theta)\sinh(\theta) R_{1413} \\ &\quad + \sinh^2(\theta) R_{1313}.\end{aligned} \tag{2.50}$$

An observer traversing the timelike geodesic which has tangent vector $\overline{E}_4$ at p has a rest space at p given by $\overline{E}_4^{\perp} = \operatorname{span}\{\overline{E}_1, \overline{E}_2, \overline{E}_3\}$. To investigate tidal accelerations for this observer, one may consider the Jacobi equation $d^2 J^i / dt^2 = -\sum_{k=1}^{3} \overline{R}_{i4k4} J^k$ using vectors in this rest space of the form $J = \sum_{i=1}^{3} J^i \overline{E}_i$ with $|J| = 1$. Notice that the tidal accelerations will be bounded as $\theta \to +\infty$ if and only if the components $\overline{R}_{i4k4}$ are bounded for large θ, and this will hold if and only if the following equations hold:

$$R_{3424} + R_{3423} = 0, \tag{2.51}$$

$$R_{3414} + R_{3413} = 0, \tag{2.52}$$

$$R_{2424} + 2R_{2423} + R_{2323} = 0, \tag{2.53}$$

$$R_{2414} + R_{2413} + R_{2314} + R_{2313} = 0, \tag{2.54}$$

$$R_{1414} + 2R_{1413} + R_{1313} = 0. \tag{2.55}$$

The vector W is nongeneric if and only if (2.53), (2.54), and (2.55) all hold, and it is nondestructive if (2.51)–(2.55) all hold [cf. Beem and Parker (1990), Beem and Harris (1993a,b)]. Tidal forces and radiation for a falling body in Schwarzschild space–time have been studied by Mashhoon (1977). Also, tidal impulses have been investigated by Mashhoon and McClune (1993). Further results on nondestructive directions have been obtained by Hall and Hossack (1993).

Remark 2.16. In order to get a physical interpretation of (2.51)–(2.55), consider a "freely falling" steel ball with center having velocity vector E_4 at p. Assume the ball has rest mass m_0 and radius $= a$. Tidal accelerations for points on the surface of the ball will correspond to $|J| = a$. Recall that the formula for the special relativistic increase in mass (or energy) is given by

$$m = \frac{m_0}{\sqrt{1 - \left(\frac{v}{c}\right)^2}} = m_0\gamma = m_0 \cosh(\theta), \tag{2.56}$$

where $\gamma = 1/\sqrt{1-(v/c)^2} = \cosh(\theta)$, v is the speed measured with respect to the original rest frame, and c is the speed of light. Of course, $v \to c$ corresponds to $\theta \to +\infty$. In other words, the classical special relativistic magnification factor of mass is $\gamma = \cosh(\theta)$. If one of (2.53), (2.54), or (2.55) fails to hold, then $W = E_3 + E_4$ is generic, and the increase in tidal accelerations for large θ is approximately proportional to γ^2 which is the square of the increase in the mass. For particles approaching the speed of light in a direction corresponding to a generic null direction, eventually the increase in tidal accelerations will become a bigger difficulty than the increase in mass. If (2.53)–(2.55) all hold, but at least one of (2.51) or (2.52) fails to hold, then the increase in tidal accelerations for large θ is approximately proportional to γ. In this case, the null direction corresponding to W is (tidally) destructive but is not generic.

2.6 The Einstein Equations

In this section we give a brief description of the Einstein equations. A heuristic derivation of these equations may be found in Frankel (1979, Chapter 3). Since these equations apply to manifolds of dimension four, we restrict our attention in this section to this dimension. The Einstein equations relate purely geometric quantities to the energy-momentum tensor T, which is a physical quantity. They may thus be used to state energy conditions in terms of T. In the case of a perfect fluid, the energy-momentum tensor also takes a simple form. This is important in general relativity because the matter of the universe is assumed to behave like a perfect fluid in the standard cosmological models. The physical motivation for studying Lorentzian manifolds is the assumption that a gravitational field may be effectively modeled by some Lorentzian metric g defined on a suitable four-dimensional manifold M. Since every manifold which admits a Lorentzian metric clearly admits uncountably many such metrics, it is necessary to decide both which manifold and which Lorentzian metric on that manifold should be used to model a given gravitational problem. The Einstein equations relate the metric tensor g, Ricci curvature Ric, and scalar curvature τ $(= R)$ to the energy-momentum tensor T. The tensor T is to be determined from physical considerations dealing with the distribution of matter and energy [cf. Hawking and Ellis (1973, Chapter 3), Misner, Thorne, and Wheeler (1973, Chapter 5)]. The *Einstein equations* may be written invariantly as

$$(2.57) \qquad \mathrm{Ric} - \frac{1}{2} R\, g + \Lambda\, g = 8\pi T \qquad (\textit{Einstein equations})$$

where Λ is a constant known as the *cosmological constant*. The constant factor of 8π is present for scaling purposes. In local coordinates, one has

$$(2.58) \quad R_{ij} - \frac{1}{2} R\, g_{ij} + \Lambda\, g_{ij} = 8\pi T_{ij} \qquad (\textit{Einstein equations in coordinates})$$

where $1 \le i, j \le 4$. The Ricci curvature and scalar curvature involve the first and second partial derivatives of the components g_{ij} of the metric g but do not involve any higher derivatives. Hence, the Einstein equations represent (nonlinear) partial differential equations in the metric and its first two derivatives. These sixteen equations reduce to ten because all of the tensors in equation (2.58) are symmetric. There is a further reduction to six equations [cf. Misner, Thorne, and Wheeler (1973, p. 409)] using the curvature identity

$$\sum_{j=1}^{4} \left(R^{ij} - \frac{1}{2} R\, g^{ij} + \Lambda\, g^{ij} \right)_{;j} = 0$$

which yields four conservation laws given by

$$\sum_{j=1}^{4} T^{ij}{}_{;j} = 0. \qquad (\textit{conservation laws})$$

Here, "$;j$" denotes covariant differentiation in the x^j direction (i.e., $\nabla_{\partial/\partial x^j}$). The Einstein equations do not determine the metric on M without sufficient boundary conditions. For example, let $M = \{(t,r) \mid t \in \mathbb{R} \text{ and } r > 2m\} \times S^2$. Then M is topologically $\mathbb{R}^2 \times S^2$. Let $\Lambda = 0$ and $T = 0$, and set $d\Omega^2 = d\theta^2 + \sin^2(\theta) d\phi^2$. Then M with this Λ and T admits both the flat metric $ds^2 = -dt^2 + dr^2 + r^2 d\Omega^2$ as well as the *Schwarzschild metric*

$$(2.59) \qquad ds^2 = -\left(1 - \frac{2m}{r}\right) dt^2 + \frac{dr^2}{\left(1 - \frac{2m}{r}\right)} + r^2 d\Omega^2 \qquad (\textit{Schwarzschild})$$

as solutions to the Einstein equations. Each of these metrics is asymptotically flat, and each is *Ricci flat* (i.e., $\mathrm{Ric} = 0$). However, the Schwarzschild metric has a nonzero curvature tensor, and hence the two metrics cannot be isometric. Nevertheless, a counting argument shows that, in general, one expects the Einstein equations to determine the metric up to diffeomorphism [cf. Hawking and Ellis (1973, p. 74)]. First, notice that the metric tensor g has 16 components which, by symmetry, reduce to ten independent components. Furthermore, four of these ten components can be accounted for by the dimension of M which allows four degrees of freedom. Thus the metric tensor is thought of as having six independent components after symmetry and diffeomorphism freedom are taken into account. Consequently, the Einstein equations yield six independent equations to determine six essential components of the metric tensor.

More rigorous approaches to the problem of existence and uniqueness of solutions to the Einstein equations using Cauchy surfaces with initial data may be found in a number of articles and books such as Chrusciel (1991), Hawking and Ellis (1973, Chapter 7), Marsden, Ebin, and Fischer (1972, pp. 233–264), and Choquet–Bruhat and Geroch (1969).

The Einstein equations may be used to relate the *timelike convergence condition* ($\mathrm{Ric}(v,v) \geq 0$ for all timelike, and hence also all null vectors v) to the energy-momentum tensor. In order to evaluate the scalar curvature R in terms of T at $p \in M$, let $\{e_1, e_2, \ldots, e_n\}$ be an orthonormal basis of T_pM, and use equation (2.57) to obtain

$$\sum_{i=1}^{4} g(e_i,e_i)\left[\mathrm{Ric}(e_i,e_i) - \frac{1}{2}R\,g\,(e_i,e_i) + \Lambda\, g(e_i,e_i)\right] = 8\pi \sum_{i=1}^{4} g(e_i,e_i)\,T(e_i,e_i).$$

Using the fact that the scalar curvature R is the trace of the Ricci curvature, this last equation becomes

$$R - 2R + 4\Lambda = 8\pi\,\mathrm{tr}(T).$$

Hence,

$$R = -8\pi\,\mathrm{tr}(T) + 4\Lambda. \tag{2.60}$$

The Einstein equations become

$$\mathrm{Ric} - \frac{1}{2}(-8\pi\,\mathrm{tr}(T) + 4\Lambda)g + \Lambda g = 8\pi T.$$

Thus,

$$\mathrm{Ric} = 8\pi\left(T - \frac{\mathrm{tr}(T)}{2}g + \frac{\Lambda}{8\pi}g\right). \tag{2.61}$$

This last equation shows that the condition $\mathrm{Ric}(v,v) \geq 0$ is equivalent to the inequality $T(v,v) \geq [\mathrm{tr}(T)/2 - \Lambda/8\pi]\, g(v,v)$. If follows that when $\Lambda = 0$ and $\dim(M) = 4$, the condition

$$\mathrm{Ric}(v,v) \geq 0$$

is equivalent to the condition

$$T(v,v) \geq \frac{\mathrm{tr}(T)}{2}\, g(v,v)$$

[cf. Hawking and Ellis (1973, p. 95)]. Note that (2.60) and (2.61) show that if $\Lambda = 0$, then $T = 0$ (i.e., vacuum) is equivalent to $\mathrm{Ric} = 0$ (i.e., Ricci flat).

The Einstein equations are fundamental in the construction of cosmological models. Consider a fluid which moves through space. This motion generates timelike flow lines in space–time. Let v be the unit speed timelike vector field which is everywhere tangent to the flow lines of the fluid. The fluid is said to be a *perfect fluid* if it has an energy density μ, pressure p, and energy-momentum tensor T such that

(2.62) $$T = (\mu + p)\, \omega \otimes \omega + p\, g, \qquad (\textit{perfect fluid})$$

which is

$$T_{ij} = (\mu + p) v_i v_j + p\, g_{ij}$$

in local coordinates. Here $\omega = \sum v_i dx^i$ is the one-form corresponding to the vector field $v = \sum v^i \partial/\partial x^i$. It follows from the above form of T that a perfect fluid is an isotropic fluid which is free of shear and viscosity. Let (M, g) be a manifold for which T has the above perfect fluid form. If the vectors $\{e_1, e_2, e_3, e_4\}$ form an orthonormal basis for T_pM, then the trace of T may be calculated as follows:

(2.63) $$\begin{aligned} \mathrm{tr}(T) &= \sum_{i=1}^{4} g(e_i, e_i)\, T(e_i, e_i) \\ &= -(\mu + p) + 4p \\ &= 3p - \mu. \end{aligned}$$

Using equation (2.61), it follows that the timelike convergence condition for a perfect fluid is equivalent to

$$T(w,w) \geq \left(\frac{3p-\mu}{2} - \frac{\Lambda}{8\pi}\right) g(w,w)$$

for all timelike (and null) w. For Lorentzian manifolds, it is easy to verify that the inner product of a timelike vector and another timelike (or nontrivial null) vector is nonzero. Thus, we may assume without loss of generality that $g(v,w) \neq 0$. Using equation (2.62) we obtain

$$(\mu + p)[g(v,w)]^2 + p\,g(w,w) \geq \left(\frac{3p-\mu}{2} - \frac{\Lambda}{8\pi}\right) g(w,w)$$

which simplifies to

$$(\mu + p)[g(v,w)]^2 \geq \left(\frac{p-\mu}{2} - \frac{\Lambda}{8\pi}\right) g(w,w). \tag{2.64}$$

Since $g(w,w) \leq 0$ and $g(v,w) \neq 0$, equation (2.64) shows that a negative cosmological constant has the effect of making the timelike convergence condition more plausible and that a positive cosmological constant has the opposite effect. Einstein originally introduced the cosmological constant because the Einstein equations with $\Lambda = 0$ predict a universe which is either expanding or contracting, and in the early part of this century it was believed that the universe was essentially static. After the discovery that the universe was expanding, the original motivation for the cosmological constant was removed; however, removing Λ from the theory has been more difficult. While astronomical experiments have failed to detect a Λ different from zero, one may always argue that Λ is so small that the experiments have not been sufficiently sensitive.

Discussions of the experimental evidence for general relativity may be found in a number of books such as Misner, Thorne, and Wheeler (1973) and Will (1981).

CHAPTER 3

LORENTZIAN MANIFOLDS AND CAUSALITY

Sections 3.1 and 3.2 give a brief review of elementary causality theory basic to this monograph as well as to general relativity. Then Section 3.3 describes an important relationship between the limit curve topology and the C^0 topology for sequences of nonspacelike curves in strongly causal space–times. Namely, if $\gamma : [a, b] \to M$ is a future directed nonspacelike limit curve of a sequence $\{\gamma_n\}$ of future directed nonspacelike curves, then a subsequence converges to γ in the C^0 topology. This result is useful for constructing maximal geodesics in strongly causal space–times using the Lorentzian distance function (cf. Chapter 8 and Chapter 12, Section 4).

In Section 3.4 we study the causal structure of two-dimensional Lorentzian manifolds. In particular, we show that if (M, g) is a space–time homeomorphic to $\mathbb{R}^2$, then (M, g) is stably causal.

Section 3.5 gives a brief discussion of the theory of Lorentzian submanifolds and the second fundamental form needed for our discussion of singularity theory in Chapter 12.

An important splitting theorem of Geroch (1970a) guarantees that a globally hyperbolic space–time may be written as a topological (although not necessarily *metric*) product $\mathbb{R} \times S$ where S is a Cauchy hypersurface. This result suggests that product space–times of the form $(\mathbb{R} \times M, -dt^2 \oplus g)$ with (M, g) a Riemannian manifold should be studied. While this class of space–times includes Minkowski space and the Einstein static universe, it fails to include the physically important exterior Schwarzschild and Robertson–Walker solutions to Einstein's equations.

In Sections 3.6 and 3.7 we study a more general class of product space–times, the so-called *warped products*, which are space–times $M_1 \times_f M_2$ with metrics of the form $g_1 \oplus f g_2$. This class of metrics, studied for Riemannian manifolds by Bishop and O'Neill (1969) and later for semi-Riemannian manifolds by O'Neill (1983), includes products, the exterior Schwarzschild space–times, and Robertson–Walker space–times. The following result, which may be regarded as a "metric converse" to Geroch's splitting theorem, is typical of the results of Section 3.6. Let $(\mathbb{R} \times M, -dt^2 \oplus g)$ be a Lorentzian product manifold with (M, g) an arbitrary Riemannian manifold. Then the following are equivalent:

(1) (M, g) is a complete Riemannian manifold.
(2) $(\mathbb{R} \times M, -dt^2 \oplus g)$ is globally hyperbolic.
(3) $(\mathbb{R} \times M, -dt^2 \oplus g)$ is geodesically complete.

3.1 Lorentzian Manifolds and Convex Normal Neighborhoods

Let M be a smooth connected paracompact Hausdorff manifold, and let TM denote the tangent bundle of M with $\pi : TM \to M$ the usual bundle map taking each tangent vector to its base point. Recall that a *Lorentzian metric* g for M is a smooth symmetric tensor field of type $(0, 2)$ on M such that for each $p \in M$, the tensor $g|_p : T_pM \times T_pM \to \mathbb{R}$ is a nondegenerate inner product of signature $(1, n-1)$ [i.e., $(-, +, \cdots, +)$]. *All noncompact manifolds admit Lorentzian metrics. However, a compact manifold admits a Lorentzian metric if and only if its Euler characteristic vanishes* [cf. Steenrod (1951, p. 207)]. The space of all Lorentzian metrics for M will be denoted by $\text{Lor}(M)$.

A continuous vector field X on M is *timelike* if $g(X(p), X(p)) < 0$ for all points $p \in M$. In general, a Lorentzian manifold does not necessarily have globally defined timelike vector fields. If (M, g) does admit a timelike vector field $X \in \mathfrak{X}(M)$, then (M, g) is said to be *time oriented* by X. The timelike vector field X divides all nonspacelike tangent vectors into two separate classes, called future and past directed. A nonspacelike tangent vector $v \in T_pM$ is said to be *future* [respectively, *past*] *directed* if $g(X(p), v) < 0$ [respectively, $g(X(p), v) > 0$]. A Lorentzian manifold (M, g) is said to be *time orientable* if (M, g) admits a time orientation by some timelike vector field X. In this case, (M, g) admits two distinct time orientations defined by X and $-X$, respectively. A time oriented Lorentzian manifold is called a space–time.

More precisely, we have the following definition.

Definition 3.1. (*Space–time*) A *space-time* (M, g) is a connected C^∞ Hausdorff manifold of dimension two or greater which has a countable basis, a Lorentzian metric g of signature $(-, +, \cdots, +)$, and a time orientation.

We now show how to construct a time oriented two-sheeted Lorentzian covering manifold $\pi : (\widetilde{M}, \widetilde{g}) \to (M, g)$ for any Lorentzian manifold (M, g) which is not time orientable.

To this end, first let (M, g) be an arbitrary Lorentzian manifold. Fix a base point $p_0 \in M$. Give a time orientation to $T_{p_0}M$ by choosing a timelike tangent vector $v_0 \in T_{p_0}M$ and defining a nonspacelike $w \in T_{p_0}M$ to be future [respectively, past] directed if $g(v_0, w) < 0$ [respectively, $g(v_0, w) > 0$]. Now let q be any point of M. Piecewise smooth curves $\gamma : [0, 1] \to M$ with $\gamma(0) = p_0$ and $\gamma(1) = q$ may be divided into two equivalence classes as follows. Given $\gamma_1, \gamma_2 : [0, 1] \to M$ with $\gamma_1(0) = \gamma_2(0) = p_0$ and $\gamma_1(1) = \gamma_2(1) = q$, let V_1 (respectively, V_2) be the unique parallel field along γ_1 (respectively, γ_2) with $V_1(0) = V_2(0) = v_0$. We say that γ_1 and γ_2 are equivalent if $g(V_1(1), V_2(1)) < 0$. If γ_1 and γ_2 are homotopic curves from p_0 to q, then γ_1 and γ_2 are equivalent. But equivalent curves are not necessarily homotopic.

Given $\gamma : [0, 1] \to M$ with $\gamma(0) = p_0$, let $[\gamma]$ denote the equivalence class of γ. Let $\widetilde{M}$ consist of all such equivalence classes of piecewise smooth curves $\gamma : [0, 1] \to M$ with $\gamma(0) = p_0$. Define $\pi : \widetilde{M} \to M$ by $\pi([\gamma]) = \gamma(1)$. If (M, g) is time orientable, then $\widetilde{M} = M$. Otherwise, $\pi : \widetilde{M} \to M$ is a two-sheeted covering [cf. Markus (1955, p. 412)].

Suppose now that the Lorentzian manifold (M, g) is *not* time orientable. It is standard from covering space theory to give the set $\widetilde{M}$ a topology and differentiable structure such that $\pi : \widetilde{M} \to M$ is a two-sheeted covering manifold. Define a Lorentzian metric $\widetilde{g}$ for $\widetilde{M}$ by $\widetilde{g} = \pi^* g$, i.e., $\widetilde{g}(v, w) = g(\pi_* v, \pi_* w)$. Then the map $\pi : \widetilde{M} \to M$ is a local isometry.

In order to show that $(\widetilde{M}, \widetilde{g})$ is time orientable, it is useful to establish a preliminary lemma. Fix a base point $\widetilde{p}_0 \in \pi^{-1}(p_0)$ for $\widetilde{M}$. Let $\widetilde{v}_0 \in T_{\widetilde{p}_0}\widetilde{M}$ be the unique timelike tangent vector in $T_{\widetilde{p}_0}\widetilde{M}$ with $\pi_* \widetilde{v}_0 = v_0$.

Lemma 3.2. *Let $\widetilde{q} \in \widetilde{M}$ be arbitrary and let $\widetilde{\gamma}_1, \widetilde{\gamma}_2 : [0,1] \to \widetilde{M}$ be two piecewise smooth curves with $\widetilde{\gamma}_1(0) = \widetilde{\gamma}_2(0) = \widetilde{p}_0$ and $\widetilde{\gamma}_1(1) = \widetilde{\gamma}_2(1) = \widetilde{q}$. If $\widetilde{V}_1, \widetilde{V}_2$ are the parallel vector fields along $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$, respectively, with $\widetilde{V}_1(0) = \widetilde{V}_2(0) = \widetilde{v}_0$, then $\widetilde{g}(\widetilde{V}_1(1), \widetilde{V}_2(1)) < 0$.*

Proof. Let $\gamma_1 = \pi \circ \widetilde{\gamma}_1$ and $\gamma_2 = \pi \circ \widetilde{\gamma}_2$. Since $\pi : (\widetilde{M}, \widetilde{g}) \to (M, g)$ is a local isometry, the vector fields $V_1 = \pi_*(\widetilde{V}_1)$ and $V_2 = \pi_*(\widetilde{V}_2)$ are parallel fields along γ_1 and γ_2, respectively, with $V_1(0) = V_2(0) = \pi_*\widetilde{v}_0 = v_0$. Also, $g(V_1(1), V_2(1)) = g(\pi_*\widetilde{V}_1(1), \pi_*\widetilde{V}_2(1)) = \widetilde{g}(\widetilde{V}_1(1), \widetilde{V}_2(1))$.

Suppose now that $\widetilde{g}(\widetilde{V}_1(1), \widetilde{V}_2(1)) \not< 0$. Since $\widetilde{V}_1(1)$ and $\widetilde{V}_2(1)$ are timelike tangent vectors, it follows that $\widetilde{g}(\widetilde{V}_1(1), \widetilde{V}_2(1)) > 0$. Thus $g(V_1(1), V_2(1)) > 0$ at $q = \pi(\widetilde{q})$. By definition of the equivalence relation on piecewise smooth curves from p_0 to q, we have $[\gamma_1] \neq [\gamma_2]$. From the construction of M, we know that $\widetilde{\gamma}_1(1) = [\gamma_1]$ and $\widetilde{\gamma}_2(1) = [\gamma_2]$. Thus $\widetilde{\gamma}_1(1) \neq \widetilde{\gamma}_2(1)$, in contradiction. □

Theorem 3.3. *Suppose that (M, g) is not time orientable. Then the two-sheeted Lorentzian covering manifold $(\widetilde{M}, \widetilde{g})$ of (M, g) constructed above is time orientable and hence is a space–time.*

Proof. Let $\widetilde{p}_0$ and $\widetilde{v}_0$ be as above. Given any $\widetilde{q} \in \widetilde{M}$, let $\sigma : [0,1] \to \widetilde{M}$ be a smooth curve with $\sigma(0) = \widetilde{p}_0$, $\sigma(1) = \widetilde{q}$. Let $\widetilde{V}$ be the unique parallel vector field along σ with $\widetilde{V}(0) = \widetilde{v}_0$. Set $F^+(\widetilde{q}) = \{\text{timelike } w \in T_{\widetilde{q}}\widetilde{M} : \widetilde{g}(\widetilde{V}(1), w) < 0\}$. By Lemma 3.2, the definition of $F^+(\widetilde{q})$ is independent of the choice of σ. Hence $\widetilde{q} \to F^+(\widetilde{q})$ consistently assigns a future cone to each tangent space $T_{\widetilde{q}}\widetilde{M}$, $\widetilde{q} \in \widetilde{M}$.

Now let h be an auxiliary positive definite Riemannian metric for $\widetilde{M}$. We may define a continuous nowhere zero timelike vector field X on $\widetilde{M}$ by choosing $X(\widetilde{q})$ to be the vector in $F^+(\widetilde{q})$ which is the unique h-unit vector in $F^+(\widetilde{q})$ having a negative eigenvalue of $\widetilde{g}$ with respect to h. That is, we may find a continuous function $\lambda : \widetilde{M} \to (-\infty, 0)$ and a continuous timelike vector field X on $\widetilde{M}$ satisfying $X(\widetilde{q}) \in F^+(\widetilde{q})$, $h(X(\widetilde{q}), X(\widetilde{q})) = 1$, and $\widetilde{g}(X(\widetilde{q}), v) = \lambda(\widetilde{q})h(X(\widetilde{q}), v)$ for all $v \in T_{\widetilde{q}}\widetilde{M}$ and $\widetilde{q} \in \widetilde{M}$. □

Implicit in the proof of Theorem 3.3 is an alternative definition for the time orientability of a Lorentzian manifold (M, g). Namely, (M, g) is time orientable if, fixing any base point $p_0 \in M$ and timelike tangent vector $v_0 \in T_{p_0}M$, the following condition is satisfied for all $q \in M$. Let $\gamma_1, \gamma_2 : [0,1] \to M$ be any two smooth curves from p to q. If V_i is the unique parallel vector field along γ_i with $V_i(0) = v_0$ for $i = 1, 2$, then $g(V_1(1), V_2(1)) < 0$. This condition means that parallel translation of the future cone determined by v_0 at p_0 to any other point q of M is independent of the choice of path from p to q. Hence a consistent choice of future timelike vectors for each tangent space may be made by parallel translation from p_0.

Recall that a smooth curve in (M, g) is said to be *timelike* (respectively, *nonspacelike, null, spacelike*) if its tangent vector is always timelike (respectively, nonspacelike, null, spacelike). As in the Riemannian case, a *geodesic* $c : (a, b) \to M$ is a smooth curve whose tangent vector moves by parallel displacement, i.e., $\nabla_{c'} c'(t) = 0$ for all $t \in (a, b)$. The tangent vector field $c'(t)$ of a geodesic c satisfies $g(c'(t), c'(t)) = \text{constant}$ for all $t \in (a, b)$ since

$$\frac{d}{dt}[g(c'(t), c'(t))] = c'(t)[g(c'(t), c'(t))] = 2\,g(\nabla_{c'} c'(t), c'(t)) = 0.$$

Consequently, a geodesic which is timelike (respectively, null, spacelike) for some value of its parameter is timelike (respectively, null, spacelike) for all values of its parameter.

The *exponential map* $\exp_p : T_pM \to M$ is defined for Lorentzian manifolds just as for Riemannian manifolds. Given $v \in T_pM$, let $c_v(t)$ denote the unique geodesic in M with $c_v(0) = p$ and $c_v{}'(0) = v$. Then the exponential $\exp_p(v)$ of v is given by $\exp_p(v) = c_v(1)$ provided $c_v(1)$ is defined.

Let $v_1, v_2, \ldots, v_n$ be any basis for the tangent space T_pM. For sufficiently small $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, the map

$$x_1v_1 + x_2v_2 + \cdots + x_nv_n \to \exp_p(x_1v_1 + x_2v_2 + \cdots + x_nv_n)$$

is a diffeomorphism of a neighborhood of the origin of T_pM onto a neighborhood $U(p)$ of p in M. Thus, assigning coordinates $(x_1, x_2, \ldots, x_n)$ to the point $\exp_p(x_1v_1 + x_2v_2 + \cdots + x_nv_n)$ in $U(p)$ defines a coordinate chart for M called *normal coordinates based at p for $U(p)$*. The set $U(p)$ is said to be a (*simple*) *convex neighborhood* of p if any two points in $U(p)$ can be joined by a unique geodesic segment of (M, g) lying entirely in $U(p)$. Whitehead (1932) has shown that any semi-Riemannian (hence Lorentzian) manifold has convex neighborhoods about each point [cf. Hicks (1965, pp. 133–136)]. In fact, it may even be assumed that for each $q \in U(p)$, there are normal coordinates based at q containing $U(p)$. We call such a neighborhood $U(p)$ a *convex normal neighborhood* [cf. Hawking and Ellis (1973, p. 34)].

The next proposition is essential to the study of the local behavior of causality. A proof is given in Hawking and Ellis (1973, pp. 103–105). It is important to carefully note the phrase "contained in U" in the statement of this result.

Proposition 3.4. *Let U be a convex normal neighborhood of q. Then the points of U which can be reached by timelike (respectively, nonspacelike) curves contained in U are those of the form $\exp_q(v)$, $v \in T_qM$, such that $g(v,v) < 0$ (respectively, $g(v,v) \leq 0$).*

3.2 Causality Theory of Space–times

In a space–time (M, g) a (nowhere vanishing) nonspacelike vector field along a curve cannot continuously change from being future directed to being past directed. It follows that a smooth timelike, null, or nonspacelike curve in (M, g) is either always future directed or always past directed.

We will use the standard notation $p \ll q$ if there is a smooth future directed timelike curve from p to q, and $p \leqslant q$ if either $p = q$ or there is a smooth future directed nonspacelike curve from p to q. Furthermore, $p < q$ will mean $p \leqslant q$ and $p \neq q$.

A continuous curve $\gamma : (a, b) \to M$ is said to be a future directed nonspacelike curve if for each $t_0 \in (a, b)$ there is an $\epsilon > 0$ and a convex normal neighborhood $U(\gamma(t_0))$ of $\gamma(t_0)$ with $\gamma(t_0 - \epsilon, t_0 + \epsilon) \subseteq U(\gamma(t_0))$ such that given any t_1, t_2 with $t_0 - \epsilon < t_1 < t_2 < t_0 + \epsilon$, there is a smooth future directed nonspacelike curve in $\big(U(\gamma(t_0)), g \,|\, U(\gamma(t_0))\big)$ from $\gamma(t_1)$ to $\gamma(t_2)$. It is necessary to use the convex normal neighborhood $U(\gamma(t_0))$ in this definition for the following reason. There exist space–times for which $p \ll q$ for all $(p, q) \in M \times M$. But in these space–times, any continuous curve γ satisfies both $\gamma(t_1) \ll \gamma(t_2)$ and $\gamma(t_2) \ll \gamma(t_1)$ for all t_1 and t_2 in the domain of γ.

For a given $p \in M$, the *chronological future* $I^+(p)$, *chronological past* $I^-(p)$, *causal future* $J^+(p)$, and *causal past* $J^-(p)$ of p are defined as follows:

$$\begin{aligned} I^+(p) &= \{q \in M \ : \ p \ll q\}, && (\textit{chronological future}) \\ I^-(p) &= \{q \in M \ : \ q \ll p\}, && (\textit{chronological past}) \\ J^+(p) &= \{q \in M \ : \ p \leqslant q\}, && (\textit{causal future}) \\ J^-(p) &= \{q \in M \ : \ q \leqslant p\}. && (\textit{causal past}) \end{aligned}$$

For general subsets $S \subseteq M$, the sets $I^+(S)$, $I^-(S)$, $J^+(S)$, and $J^-(S)$ are defined analogously: for example, $I^+(S) = \{q \in M \mid s \ll q \text{ for some } s \in S\}$.

The relations $\ll$ and $\leqslant$ are clearly transitive. Moreover,

$$p \ll q \quad \text{and} \quad q \leqslant r \quad \text{implies} \quad p \ll r,$$

and

$$p \leqslant q \quad \text{and} \quad q \ll r \quad \text{implies} \quad p \ll r$$

[cf. Penrose (1972, p. 14)]. If there is a future directed timelike curve from p to q, there is a neighborhood U of q such that any point of U can be reached by a future directed timelike curve. Consequently, it follows that

Lemma 3.5. *If p is any point of the space–time (M, g), then $I^+(p)$ and $I^-(p)$ are open sets of M.*

An example has been given in Chapter 1, Figure 1.1, to show that the sets $J^+(p)$ and $J^-(p)$ are neither open nor closed in general.

Two especially important classes of subsets of a space–time are future sets and past sets. In this section we will restrict our attention to *open* future and past sets, which may be defined as follows.

Definition 3.6. (*Future and Past Sets*) The (open) subset F (respectively, P) of the space–time (M, g) is said to be a *future* (respectively, *past*) set if $F = I^+(F)$ (respectively, $P = I^-(P)$).

These sets will be used in studying the causality of the gravitational plane wave space–times in Chapter 13. Future and past sets have often proven useful in causality proofs [cf. Hawking and Sachs (1974), Dieckmann (1987)]. Rather complete discussions of this topic may be found in Penrose (1972, Section 3) and in Hawking and Ellis (1973).

The simplest examples are formed by taking $F = I^+(S)$ or $P = I^-(S)$, where S is an arbitrary subset of M. (Indeed, these sets may also be defined in this manner.) They also share the following property with the chronological past and future sets $I^-(p)$ and $I^+(p)$ of an arbitrary point p in M:

$$x \text{ in } F \text{ and } x \ll y \text{ implies } y \text{ in } F; \qquad (\textit{future set})$$

$$x \text{ in } P \text{ and } y \ll x \text{ implies } y \text{ in } P. \qquad (\textit{past set}).$$

The next two results give simple characterizations of the closures and boundaries of sets which are either future or past.

Proposition 3.7.

(1) *If F is a future set, then $\overline{F} = \{x \in M : I^+(x) \subseteq F\}$.*
(2) *If P is a past set, then $\overline{P} = \{x \in M : I^-(x) \subseteq P\}$.*

Proof. As usual, it suffices to establish (1). First, suppose $I^+(x) \subseteq F$. We may take $\{q_n\} \subseteq I^+(x)$ with $\lim_{n\to\infty} q_n = x$. Since $\{q_n\} \subseteq F$, this yields $x \in \overline{F}$. Conversely, let $x \in \overline{F}$ be given. Select any $z \in I^+(x)$. Then $x \in I^-(z)$ which is open, and since $x \in \overline{F}$, there exists $y \in I^-(z) \cap F$. But then $y \in F$ and $y \ll z$ implies $z \in F$ since F is a future set. Thus $I^+(x) \subseteq F$ as required. □

Corollary 3.8. *Let F (respectively, P) be a future (respectively, past) set. Then*

(1) $\partial(F) = \{x \in M : x \notin F$ *and* $I^+(x) \subseteq F\}$, *and*
(2) $\partial(P) = \{x \in M : x \notin P$ *and* $I^-(x) \subseteq P\}$.

In the particular cases where $F = I^+(p)$ or $P = I^-(p)$, the following useful characterizations of

$$\overline{I^-(p)} = \overline{J^-(p)} \qquad \text{and} \qquad \overline{I^+(p)} = \overline{J^+(p)}$$

are obtained for arbitrary space–times:

$$\overline{I^+(p)} = \{x \in M : I^+(x) \subseteq I^+(p)\}, \qquad (closure)$$
$$\overline{I^-(p)} = \{x \in M : I^-(x) \subseteq I^-(p)\},$$
$$\partial(I^+(p)) = \{x \in M : x \notin I^+(p) \text{ and } I^+(x) \subseteq I^+(p)\}, \qquad (boundary)$$
$$\text{and} \quad \partial(I^-(p)) = \{x \in M : x \notin I^-(p) \text{ and } I^-(x) \subseteq I^-(p)\}.$$

Past and future sets have played an important role in singularity theory in general relativity as treated in Penrose (1972) or Hawking and Ellis (1973) via the allied concept of achronal boundaries. A subset B of M is said to be an *achronal boundary* if $B = \partial(F)$ for some future set F. (Of course, it is necessary to prove that $\partial(F)$ is achronal, i.e., that no two points of $\partial(F)$ may be joined by a future timelike curve, for this definition to make sense.) In particular, for any p in M the set

$$\partial(I^+(p)) = \partial(J^+(p))$$

is a simple example of an achronal boundary. Achronal boundaries have the following important regularity properties.

Theorem 3.9. *The (nonempty) boundary $\partial(F)$ of a future set is a closed achronal (Lipschitz) topological hypersurface.*

Discussions of this result are given in Penrose (1972, pp. 21–23) and in Hawking and Ellis (1973, p. 187) from the viewpoint of applications in singularity theory. A rather complete treatment of a number of mathematical aspects may be found in O'Neill (1983, pp. 413–415).

It may happen that $p \in I^+(p)$. If so, there is a closed timelike curve through p, and the space–time is said to have a causality violation. For example, on the cylinder $M = S^1 \times \mathbb{R}$ with the Lorentzian metric $ds^2 = -d\theta^2 + dt^2$, the circles $t =$ constant are closed timelike curves. In this space–time, $I^+(p) = M$ for all $p \in M$. A number of causality conditions have been defined in general relativity in recent years because of the problems associated with examples of causality violations.

Space–times which do not contain any closed timelike curves [i.e., $p \notin I^+(p)$ for all $p \in M$] are said to be *chronological*. A space–time with no closed nonspacelike curves is said to be *causal*. Equivalently, a causal space–time contains no pair of distinct points p and q with $p \leqslant q \leqslant p$. The cylinder $M = S^1 \times \mathbb{R}$ with the Lorentzian metric $ds^2 = d\theta\, dt$ is an example of a chronological space–time that fails to be causal. The only closed nonspacelike curves in this example are the circles $t =$ constant, which with proper parametrization are null geodesics.

The chronological condition is the weakest causality condition which will be introduced. The next proposition guarantees that no compact space–time is either causal or chronological.

Proposition 3.10. *Any compact space–time (M, g) contains a closed timelike curve and thus fails to be chronological.*

Proof. Since the sets of the form $I^+(p)$ are open, it may be seen that $\{I^+(p) : p \in M\}$ forms an open cover of M. By compactness, we may extract a finite subcover $\{I^+(p_1), I^+(p_2), \ldots, I^+(p_k)\}$. Now $p_1 \in I^+(p_{i(1)})$ for some $i(1)$ with $1 \leq i(1) \leq k$. Similarly, $p_{i(1)} \in I^+(p_{i(2)})$ for some index $i(2)$ with $1 \leq i(2) \leq k$. Continuing inductively, we obtain an infinite sequence $\cdots \ll p_{i(3)} \ll p_{i(2)} \ll p_{i(1)} \ll p_1$. Since k is finite, there are only a finite number of distinct $p_{i(j)}$'s. Thus there are repetitions on the list, and from the transitivity of $\ll$, it follows that $p_{i(n)} \in I^+(p_{i(n)})$ for some index $p_{i(n)}$. Thus (M, g) contains a closed timelike curve through $p_{i(n)}$. $\square$

Tipler (1979) has proved that certain classes of compact space–times contain closed timelike *geodesics*, not just closed timelike curves. Since the proof uses the Lorentzian distance function as a tool, discussion of Tipler's result is postponed until Section 4.1, Theorem 4.15.

A space–time is said to be *distinguishing* if for all points p and q in M, either $I^+(p) = I^+(q)$ or $I^-(p) = I^-(q)$ implies $p = q$. In a distinguishing space–time, distinct points have distinct chronological futures and chronological pasts. Thus, points are distinguished both by their chronological futures and pasts.

A distinguishing space–time is said to be *causally continuous* if the set-valued functions I^+ and I^- are outer continuous. Since I^+ and I^- are always inner continuous [cf. Hawking and Sachs (1974, p. 291)], the causally continuous space–times are those distinguishing space–times for which both the chronological future and past of a point vary continuously with the point. Here I^+ is said to be *inner continuous* at $p \in M$ if for each compact set $K \subseteq I^+(p)$, there exists a neighborhood $U(p)$ of p such that $K \subseteq I^+(q)$ for each $q \in U(p)$. The set-valued function I^+ is *outer continuous* at $p \in M$ if for each compact set $K \subseteq M - \overline{I^+(p)}$, there exists some neighborhood $U(p)$ of p such that $K \subseteq M - \overline{I^+(q)}$ for each $q \in U(p)$. Inner and outer continuity of I^- may be defined dually. An example of a space–time for which I^- fails to be outer continuous is given in Figure 3.1. The concept of causal continuity was introduced by Hawking and Sachs (1974). For these space–times the causal structure may be extended to the causal boundary [cf. Budic and Sachs (1974)]. Furthermore, a metrizable topology may be defined on the causal completion of a causally continuous space–time [cf. Beem (1977)].

An open set U in a space–time is said to be *causally convex* if no nonspacelike curve intersects U in a disconnected set. Given $p \in M$, the space–time (M, g) is said to be *strongly causal at* p if p has arbitrarily small causally convex neighborhoods. Thus, p has arbitrarily small neighborhoods such that no nonspacelike curve that leaves one of these neighborhoods ever returns. The space–time (M, g) is *strongly causal* if it is strongly causal at every point. It may be shown that the set of points of an arbitrary space–time (M, g) at which (M, g) is strongly causal is an open subset of M [cf. Penrose (1972, p. 30)]. It is not hard to show that strongly causal space–times are distinguishing.

Strongly causal space–times may be characterized in terms of the Alexandrov topology for M. The *Alexandrov topology* on an arbitrary space–time (M, g) is the topology given M by taking as a basis all sets of the form $I^+(p) \cap I^-(q)$ with $p, q \in M$ (cf. Figure 1.2). The given manifold topology on M is always at least as fine as the Alexandrov topology since $I^+(p) \cap I^-(q)$ is an open set in the given topology by Lemma 3.5. The following result has been obtained by Kronheimer and Penrose [cf. Penrose (1972, p. 34)].

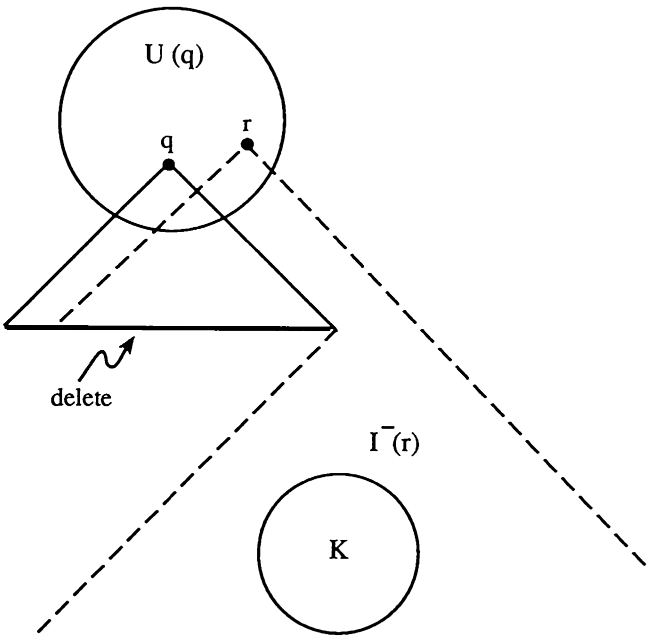


FIGURE 3.1. A space–time which is *not* causally continuous is shown. The map $p \to I^-(p)$ fails to be outer continuous at the point q. The compact set K is contained in $M - \overline{I^-(q)}$, yet each neighborhood $U(q)$ of q contains some point r such that K is not contained in $M - \overline{I^-(r)}$.

Proposition 3.11. *The Alexandrov topology for (M, g) agrees with the given manifold topology iff (M, g) is strongly causal.*

Proof. Assume first that (M, g) is strongly causal. Then each $p \in M$ has some convex normal neighborhood $U(p)$ such that no nonspacelike curve intersects $U(p)$ more than once. The set $U(p)$ is a convex normal neighborhood of each of its points, and hence Proposition 3.4 implies that for each $q \in U(p)$, the chronological future (respectively, past) of q in $\left(U(p), g|_{U(p)}\right)$ consists of all points which can be reached by geodesic segments in U of the form $\exp_q(tv)$ for $0 \leq t \leq 1$, where v is a future (respectively, past) directed timelike vector at q. This demonstrates that the Alexandrov topology on $\left(U(p), g\,|_{U(p)}\right)$ agrees with the given manifold topology on $U(p)$. Using the fact that no nonspacelike curve of (M, g) intersects $U(p)$ more than once, it follows that the Alexandrov topology agrees with the given manifold topology.

Now assume that strong causality fails to hold at $p \in M$. Then there is a convex normal neighborhood $V(p)$ of p such that if $W(p)$ is any neighborhood of p with $W(p) \subseteq V(p)$, a nonspacelike curve starts in $W(p)$, leaves $V(p)$, and returns to $W(p)$. It follows that all neighborhoods of p in the Alexandrov topology contain points outside of $V(p)$. Thus, the Alexandrov topology differs from the given manifold topology. □

In order to study causality breakdowns and geodesic incompleteness in general relativity, it is helpful to formulate the concept of *inextendibility* for nonspacelike curves. This may be done as follows. Let $\gamma : [a, b) \to M$ be a curve in M. The point $p \in M$ is said to be the *endpoint* of γ corresponding to $t = b$ if

$$\lim_{t \to b^-} \gamma(t) = p.$$

If $\gamma : [a, b) \to M$ is a future (respectively, past) directed nonspacelike curve with endpoint p corresponding to $t = b$, the point p is called a *future* (respectively, *past*) *endpoint* of γ. A nonspacelike curve is said to be *future inextendible* (or *future endless*) if it has no future endpoint. Dually, a *past inextendible* nonspacelike curve is one that has no past endpoint.

Convention 3.12. A nonspacelike curve $\gamma : (a, b) \to M$ is said to be *inextendible* (or *endless*) if it is both future and past inextendible.

Causal space–times exist that contain inextendible nonspacelike curves having compact closure. An example given by Carter is displayed in Figure 3.2 [cf. Hawking and Ellis (1973, p. 195)]. An inextendible nonspacelike curve which has compact closure and hence is contained in a compact set is said to be *imprisoned*. Thus, Carter's example shows that imprisonment can occur in causal space–times.

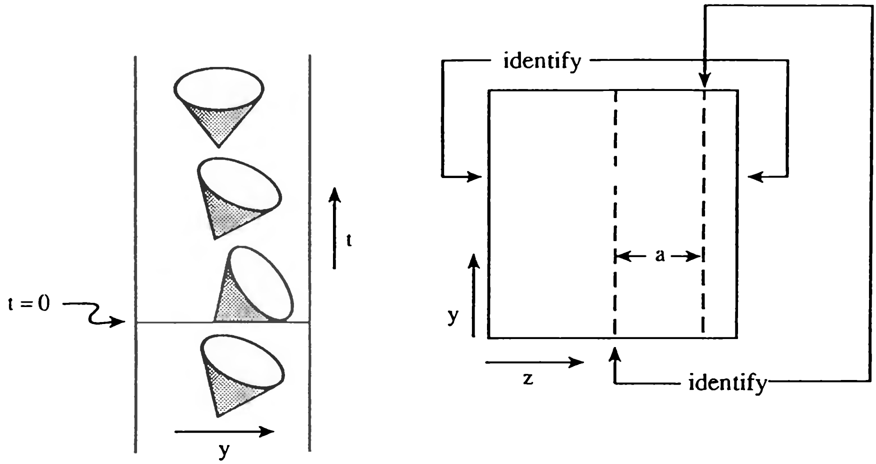


FIGURE 3.2. A causal space–time (M, g) is shown which has imprisoned nonspacelike curves that are inextendible. Let a be an irrational number, and let $M = \mathbb{R} \times S^1 \times S^1 = \{(t, y, z) \in \mathbb{R}^3 : (t, y, z) \sim (t, y, z+1) \text{ and } (t, y, z) \sim (t, y+1, z+a)\}$. The Lorentzian metric is given by $ds^2 = (\cosh t - 1)^2(dy^2 - dt^2) - dt\,dy + dz^2$.

Let $\gamma : [a, b) \to M$ be a future directed nonspacelike curve. Then γ is said to be *future imprisoned* in the compact set K if there is some $t_0 < b$ such that $\gamma(t) \in K$ for all $t_0 < t < b$. The curve γ is said to be *partially future imprisoned* in the compact set K if there exists an infinite sequence $t_n \uparrow b$ with $\gamma(t_n) \in K$ for each n.

If (M, g) is strongly causal and K is a compact subset of M, then K may be covered with a finite number of convex normal neighborhoods $\{U_i\}$ such that no nonspacelike curve which leaves some U_i ever returns to that U_i. This observation leads to the following result.

Proposition 3.13. *If (M, g) is strongly causal, then no inextendible non-spacelike curve can be partially future (or past) imprisoned in any compact set.*

We now discuss another important class of space–times in general relativity, stably causal space–times. For this purpose as well as for later use, it is helpful to define the fine C^r topologies on $\mathrm{Lor}(M)$.

Recall that $\mathrm{Lor}(M)$ denotes the space of all Lorentzian metrics on M. The *fine C^r topologies* on $\mathrm{Lor}(M)$ may be defined by using a fixed countable covering $\mathcal{B} = \{B_i\}$ of M by coordinate neighborhoods with the property that the closure of each B_i lies in a coordinate chart of M and such that each compact subset of M intersects only finitely many of the B_i's. This last requirement is the condition that the covering be locally finite. Let $\delta : M \to (0, \infty)$ be a continuous function. Then $g_1, g_2 \in \mathrm{Lor}(M)$ are said to be $\delta : M \to (0, \infty)$ close in the C^r topology, written $|g_1 - g_2|_r < \delta$, if for each $p \in M$ all of the corresponding coefficients and derivatives up to order r of the two metric tensors g_1 and g_2 are $\delta(p)$ close at p when calculated in the fixed coordinates of all $B_i \in \mathcal{B}$ which contain p. The sets $\{g_1 \in \mathrm{Lor}(M) : |g_1 - g_2|_r < \delta\}$, with $g_2 \in \mathrm{Lor}(M)$ arbitrary and $\delta : M \to (0, \infty)$ an arbitrary continuous function, form a basis for the fine C^r topology on $\mathrm{Lor}(M)$. This topology may be shown to be independent of the choice of coordinate cover $\mathcal{B}$.

The C^r topologies for $r = 0, 1, 2$ may be given the following interpretations.

Remark 3.14. (1) Two Lorentzian metrics for M which are close in the fine C^0 topology have light cones which are close.
(2) Two Lorentzian metrics for M which are close in the fine C^1 topology have geodesic systems which are close (cf. Section 7.2).
(3) Two Lorentzian metrics for M which are close in the fine C^2 topology have curvature tensors which are close.

A space–time (M, g) is said to be *stably causal* if there is a fine C^0 neighborhood $U(g)$ of g in $\mathrm{Lor}(M)$ such that each $g_1 \in U(g)$ is causal. Thus a stably causal space–time remains causal under small C^0 perturbations.

Stably causal space–times may be characterized in terms of a partial ordering $<$ for $\mathrm{Lor}(M)$ defined using light cones to compare Lorentzian metrics. Explicitly, if A is a subset of M, one defines $g_1 \leqslant_A g_2$ if for each $p \in A$ and $v \in T_pM$ with $v \neq 0$, $g_1(v,v) \leq 0$ implies $g_2(v,v) \leq 0$. One also defines $g_1 <_A g_2$ if for each $p \in A$ and $v \in T_pM$ with $v \neq 0$, $g_1(v,v) \leq 0$ implies $g_2(v,v) < 0$. We will write $g_1 < g_2$ (respectively, $g_1 \leqslant g_2$) for $g_1 <_M g_2$ (respectively, $g_1 \leqslant_M g_2$). Thus $g_1 < g_2$ means that at every point of M the light cone of g_1 is smaller than the light cone of g_2, or g_2 has wider light cones than g_1. It may be shown that (M,g) is stably causal if and only if there exists some causal $g_1 \in \mathrm{Lor}(M)$ with $g < g_1$.

A C^0 function $f : M \to \mathbb{R}$ is a *global time function* if f is strictly increasing along each future directed nonspacelike curve. A space–time (M,g) admits a global time function if and only if it is stably causal [cf. Hawking (1968), Seifert (1977)]. However, there is generally no natural choice of a time function for a stably causal space–time.

Let $f : M \to \mathbb{R}$ be a smooth function such that the gradient ∇f is always timelike. If $\gamma : (a,b) \to M$ is a future directed nonspacelike curve with nonvanishing tangent vector $\gamma'(t)$, then $g(\nabla f(\gamma(t)), \gamma'(t)) = \gamma'(t)(f)$ must either be always positive or always negative. Thus f must be either strictly increasing or strictly decreasing along γ. It follows that f must be strictly increasing or decreasing along all future directed nonspacelike curves. Hence f or $-f$ is a smooth global time function for M. Furthermore, ∇f must be orthogonal to each of the level surfaces $f^{-1}(c) = \{p \in M : f(p) = c\}$, $c \in \mathbb{R}$, of f. These level surfaces are hypersurfaces orthogonal to a timelike vector field and are *spacelike*, i.e., g restricted to each of these hypersurfaces is a positive definite metric. Since the gradient of f is nonvanishing and *df* is an exact 1-form, it follows that M is foliated by the level surfaces $\{f^{-1}(c) : c \in \mathbb{R}\}$. Each nonspacelike curve γ of M can intersect a given level surface at most once since f must be strictly increasing or decreasing along γ.

One of the most important causality conditions which we will discuss in this section is global hyperbolicity. Globally hyperbolic space–times have the important property, frequently invoked during specific geodesic constructions, that any pair of causally related points may be joined by a nonspacelike geodesic segment of maximal length.

Definition 3.15. (*Globally Hyperbolic*) A strongly causal space–time (M, g) is said to be *globally hyperbolic* if for each pair of points $p, q \in M$, the set $J^+(p) \cap J^-(q)$ is compact.

A distinguishing space–time (M, g) is *causally simple* if $J^+(p)$ and $J^-(p)$ are closed subsets of M for all $p \in M$. It then follows that

Proposition 3.16. *A globally hyperbolic space–time is causally simple.*

Proof. Suppose $q \in \overline{J^+(p)} - J^+(p)$ for some $p \in M$. Choose any $r \in I^+(q)$. Since $I^-(r)$ is open and $q \in \overline{J^+(p)}$, it may be seen that $r \in I^+(p)$ by taking a subsequence $\{q_n\} \subseteq J^+(p)$ with $q_n \to q$ and using the fact that $p \leqslant q_n$ and $q_n \ll r$ imply $p \ll r$. Consequently, $q \in \overline{J^+(p) \cap J^-(r)} - (J^+(p) \cap J^-(r))$. But this is impossible since $J^+(p) \cap J^-(r)$ is compact hence closed. □

Globally hyperbolic space–times may be characterized using Cauchy surfaces. A *Cauchy surface* S is a subset of M which every inextendible nonspacelike curve intersects exactly once. (Some authors only require that every inextendible *timelike* curve intersect S exactly once.) It may be shown that a space–time is globally hyperbolic if and only if it admits a Cauchy surface [cf. Hawking and Ellis (1973, pp. 211–212)]. Furthermore, Geroch (1970a) has established the following important structure theorem for globally hyperbolic space–times [cf. Sachs and Wu (1977b, p. 1155)].

Theorem 3.17. *If (M, g) is a globally hyperbolic space–time of dimension n, then M is homeomorphic to $\mathbb{R} \times S$ where S is an $(n-1)$-dimensional topological submanifold of M, and for each t, $\{t\} \times S$ is a Cauchy surface.*

The proof of this theorem uses a function $f : M \to \mathbb{R}$ given by $f(p) = m(J^+(p))/m(J^-(p))$ where m is a measure on M with $m(M) = 1$. The level sets of f may be seen to be Cauchy surfaces as desired, but f is not necessarily smooth.

A time function $f : M \to \mathbb{R}$ will be said to be a *Cauchy time function* if each level set $f^{-1}(c)$, $c \in \mathbb{R}$, is a Cauchy surface for M. In studying globally hyperbolic space–times, it is helpful to use Cauchy time functions rather than arbitrary time functions.

In a complete Riemannian manifold, any two points may be joined by a geodesic of minimal length. Avez (1963) and Seifert (1967) have obtained a Lorentzian analogue of this result for globally hyperbolic space–times.

Theorem 3.18. *Let (M, g) be globally hyperbolic and $p \leqslant q$. Then there is a nonspacelike geodesic from p to q whose length is greater than or equal to that of any other future directed nonspacelike curve from p to q.*

It should be emphasized that the geodesic in Theorem 3.18 is not necessarily unique. This result will also be discussed from the viewpoint of the Lorentzian distance function in Section 6.1.

It is natural to consider how continuity properties of a given candidate for a time function on a given space–time are influenced by the causal structure of the space–time. We have just indicated how Geroch (1970a) used past and future volume functions to obtain the globally hyperbolic topological splitting theorem (Theorem 3.17) and how stably causal space–times may be characterized in terms of the existence of a (continuous) global time function [cf. Hawking (1968), Seifert (1977), Hawking and Sachs (1974)]. It is often stated that an auxiliary "additive measure H on M which assigns positive volume $H[U]$ to each open set U and assigns finite volume $H[M]$ to $M \ldots$" may be employed in this context. For such a measure, it has been asserted that a distinguishing space–time (M, g) is reflecting (cf. Definition 3.20), hence causally continuous, if and only if the past and future volume functions t^- and t^+ associated to any such measure H on M are continuous global time functions.

J. Dieckmann (1987, 1988) has noted that the above assertion fails to be valid unless certain further regularity properties are imposed on the measure H. In particular, one needs to restrict attention to measures that are not positive on the boundaries of chronological futures and pasts (cf. Definition 3.19). The need for this restriction is best understood by means of an example. Let (M, g) be four-dimensional Minkowski space–time, and let m be any Borel measure for M with $m(M)$ finite and with $m\big(\partial(I^+(p))\big) = m\big(\partial(I^-(p))\big) = 0$ for all p in M. Such a measure may be constructed using a partition of unity and local volume forms as follows [cf. Hawking and Ellis (1973, p. 199), Geroch (1970a, p. 446), Dieckmann (1988, p. 860)]. Let ω be the usual volume form for all of M, and let $\{\rho_n\}$ be a partition of unity subordinate to a covering $\{U_n\}$ with each U_n being a simple region and with

$$\int_{U_n} \omega < 1.$$

Then let m be the Borel measure associated (by integration) to the 4-form

$$\sum_n 2^{-n} \rho_n \cdot \omega.$$

It is easily seen that $m\big(\partial(I^+(p))\big) = m\big(\partial(I^-(p))\big) = 0$ for any p in M.

Now define a measure H for M as follows. Put $0 = (0,0,0,0)$ and

$$H[U] = \begin{cases} m(U) & \text{if} \quad 0 \notin U, \\ m(U)+1 & \text{if} \quad 0 \in U. \end{cases}$$

Then even though (M,g) is certainly causally continuous, the past volume function $t^-(p) = H[I^-(p)]$ fails to be continuous at 0. The difficulty is that H assigns measure one to the point 0. Thus, H assigns positive measure to light cones containing this point.

This counterexample led Dieckmann (1987, 1988) to investigate more precisely, for a subclass of probability measures for a given space–time, how the past and future volume functions are related to space–time causality. We will summarize certain of these results. We begin with an arbitrary space–time (M,g) and, following Dieckmann, abstract the crucial properties of the Borel measure m whose construction was sketched above.

Definition 3.19. (*Admissible Measure*) A Borel measure m on (M,g) is said to be *admissible* provided that

(3.1) $m(U) > 0$ for all nonempty open sets U,

(3.2) $m(M) < +\infty$, and

(3.3) $m\big(\partial(I^+(p))\big) = m\big(\partial(I^-(p))\big) = 0$ for all p in M.

Clearly, the above measure H fails to satisfy condition (3.3) and thus fails to be admissible.

The past (respectively, future) volume functions (associated to the measure m) are given respectively by

$$t^-(p) = m(I^-(p)) \tag{3.4}$$

and

$$t^+(p) = -m(I^+(p)) \tag{3.5}$$

for all p in M.

The crucial importance of property (3.3) is that it implies

$$m\left(\overline{I^-(p)}\right) = m(I^-(p)) \tag{3.6}$$

and

$$m\left(\overline{I^+(p)}\right) = m(I^+(p)) \tag{3.7}$$

for all p in M.

If the given space–time happens to be *totally vicious*, so that $I^-(p) = I^+(p) = M$ for all p in M, then t^- and t^+ take on the constant values $m(M)$ and $-m(M)$, respectively. More generally, in the presence of certain causality violations the past and future volume functions are only weakly increasing along future causal curves and may also fail to be continuous. Hence the volume functions do not, in general, define generalized time functions in the sense of Definition 3.23 below without imposing some causality conditions on the space–time in question.

For convenience, we will employ the notational convention used in O'Neill (1983) that $p < q$ if there exists a (nontrivial) future directed nonspacelike curve from p to q. Thus, $p < q$ if $p \leqslant q$ and $p \neq q$. The usual transitivity relationships (i.e., $r \ll p$, $p < q$, and $q \ll s$ together imply $r \ll q$ and $p \ll s$) yield that the volume functions t^- and t^+ are (not necessarily continuous) "semi-time functions" in the sense that for any p, q in M

$$p < q \quad \text{implies} \quad t^-(p) \leq t^-(q) \quad \text{and} \quad t^+(p) \leq t^+(q). \tag{3.8}$$

It is also interesting that the usual continuity properties of a measure imply that t^- (respectively, t^+) is lower semicontinuous (respectively, upper semicontinuous).

We now introduce a causality condition important in Hawking and Sachs (1974) and in Dieckmann (1987, 1988).

Definition 3.20. (*Reflecting*) A space–time (M, g) is said to be *past reflecting* (respectively, *future reflecting*) at q in M if for all p in M

$$I^+(p) \supseteq I^+(q) \quad \text{implies} \quad I^-(p) \subseteq I^-(q), \tag{3.9}$$

respectively,

$$I^-(p) \supseteq I^-(q) \quad \text{implies} \quad I^+(p) \subseteq I^+(q) \tag{3.10}$$

and is said to be *reflecting at q* if it satisfies both conditions. The space–time is said to be *reflecting* if it is reflecting at all points.

It is well known that for space–times which fail to be reflecting, the past and future volume functions may fail to be continuous [cf. Figure 1.2 in Hawking and Sachs (1974, p. 289)]. Even more strikingly, Dieckmann obtained the following more precise relationship between continuity and reflexivity. In this result, properties (3.6) and (3.7) for the admissible measure are crucial.

Proposition 3.21. *Let (M, g) be a (not necessarily distinguishing) space–time, and let t^- and t^+ be the past and future volume functions associated to an admissible Borel measure. Then*

(1) *t^- is continuous at q iff (M, g) is past reflecting at q, and*
(2) *t^+ is continuous at q iff (M, g) is future reflecting at q.*

Moreover, from his proof of a version of Proposition 3.21, Dieckmann is able to conclude that the set of points at which the volume function t^- (respectively, t^+) fails to be continuous is a union of null geodesics without past (respectively, future) endpoints. Especially, a space–time may not fail to be reflecting at isolated points, as had already been noted in Vyas and Akolia (1986).

Thus, reflexivity settles the question of the continuity of the volume functions independent of choice of admissible measure. The "time function" aspect of being strictly increasing along future directed nonspacelike curves employs the following lemma [cf. Dieckmann (1987, p. 47)].

Lemma 3.22. *Let (M, g) be an arbitrary (not necessarily distinguishing) space–time, and let t^- be the past volume function associated to an admissible Borel measure. Suppose that p, q in M satisfy $p < q$ and $I^-(p) \neq I^-(q)$. Then*

$$t^-(p) < t^-(q). \tag{3.11}$$

Lemma 3.22 has the immediate consequence that (M, g) is chronological if and only if some past (or future) volume function is strictly increasing along all future timelike curves. More importantly, with Lemma 3.22 in hand the following proposition may now be obtained with the help of future set techniques. For clarity, let us first make precise the notion of "generalized time function" as employed in the present context.

Definition 3.23. (*Generalized Time Function*) A (not necessarily continuous) function $t : (M, g) \to \mathbb{R}$ is said to be a *generalized time function* if, for all p, q in M,

$$p < q \quad \text{implies} \quad t(p) < t(q). \tag{3.12}$$

Thus t is strictly increasing along all future nonspacelike curves but is not required to be continuous.

Proposition 3.24. *Let t^- (respectively, t^+) be the past (respectively, future) volume function on (M, g) associated to an admissible Borel measure. Then*

(1) *(M, g) is past distinguishing iff t^- is a generalized time function; and*
(2) *(M, g) is future distinguishing iff t^+ is a generalized time function.*

Now it has been established that "causal continuity" (i.e., (M, g) is distinguishing and I^+ and I^- are outer continuous) is equivalent to "reflecting" and "distinguishing" [cf. Hawking and Sachs (1974, p. 292)]. Hence, Propositions 3.21 and 3.24 imply the following characterization of causal continuity in terms of volume functions.

Theorem 3.25. *The following are equivalent:*

(1) *The space–time (M, g) is causally continuous.*

(2) *For any (and hence all) admissible Borel measures, the associated volume functions t^- and t^+ are both continuous time functions.*

The condition of "causal continuity" implies the "stable causality" condition but not conversely. Stable causality has the characterization that (M, g) admits *some* continuous global time function (*which will not, in general, be a volume function associated to some admissible Borel measure*).

We now state the characterization of global hyperbolicity in terms of volume functions [implicit in Geroch (1970a)] which is given in Dieckmann (1987).

Theorem 3.26. *Let t^- (respectively, t^+) be the past (respectively, future) volume functions associated to an admissible Borel measure m for the space–time (M, g). Define $t : (M, g) \to \mathbb{R}$ by*

$$t(p) = \ln\left(-\frac{t^-(p)}{t^+(p)}\right) \tag{3.13}$$

for p in M. Then the following are equivalent:

(1) *The space–time (M, g) is globally hyperbolic.*

(2) *The past and future volume functions t^-, t^+ are continuous time functions, and for any inextendible future directed nonspacelike curve $\gamma : (a, b) \to (M, g)$ we have $\lim_{u\to b} t^+(\gamma(u)) = \lim_{u\to a} t^-(\gamma(u)) = 0$.*

(3) *The function t given by (3.13) is a continuous time function, and for any inextendible future directed nonspacelike curve γ, $\mathrm{range}(t\circ\gamma) = \mathbb{R}$.*

(4) *For all a in $\mathbb{R}$, the set $t^{-1}(\{a\})$ is a Cauchy surface.*

The next proposition gives sufficient conditions for a space–time to be causal in terms of the volume functions [cf. Dieckmann (1987)].

Proposition 3.27. *Let t be a past or future volume function associated to an admissible Borel measure for the given space–time (M, g). Suppose further that*

(1) *$p \ll q$ implies $t(p) < t(q)$ for all p, q in M (which implies chronology), and*

(2) *for each complete null geodesic* $\beta : \mathbb{R} \to (M, g)$, *there exist* u_1, u_2 *in* $\mathbb{R}$ *such that* $t(\beta(u_1)) < t(\beta(u_2))$.

Then (M, g) *is causal.*

In Vyas and Joshi (1983), a discussion is given of how causal functions similar to (3.5) may be related to the ideal boundary points and singularities of a space–time. An interesting list of 71 assertions on causality to be proved or disproved (together with answers) has been given in Geroch and Horowitz (1979, pp. 289–293). We now give a diagram (Figure 3.3) indicating the relations between the causality conditions discussed above [cf. Hawking and Sachs (1974, p. 295), Carter (1971a)].

3.3 Limit Curves and the C^0 Topology on Curves

Two different forms of convergence for a sequence of nonspacelike curves $\{\gamma_n\}$ have been useful in Lorentzian geometry and general relativity [cf. Penrose (1972), Hawking and Ellis (1973)]. The first type of convergence uses the concept of a limit curve of a sequence of curves, while the second type uses the C^0 topology on curves. For arbitrary space–times, neither of these types of convergence is stronger than the other. However, we will show that for strongly causal space–times, these two forms of convergence are closely related. This relationship will be useful in constructing maximal geodesics in strongly causal space–times (cf. Sections 8.1 and 8.2).

Definition 3.28. (*Limit Curve*) A curve γ is a *limit curve of the sequence* $\{\gamma_n\}$ if there is a subsequence $\{\gamma_m\}$ such that for all p in the image of γ, each neighborhood of p intersects all but a finite number of curves of the subsequence $\{\gamma_m\}$. The subsequence $\{\gamma_m\}$ is said to *distinguish* the limit curve γ.

In general, a sequence of curves $\{\gamma_n\}$ may have no limit curves or may have many limit curves. This is true even if the curves $\{\gamma_n\}$ are nonspacelike. Furthermore, even in causal space–times a limit curve of a sequence of nonspacelike limit curves is not necessarily nonspacelike. For example, the curve $\gamma(u) = (0, 0, u)$ in Carter's example (cf. Figure 3.2) is not nonspacelike although it is a limit curve of any sequence $\{\gamma_n\}$ of inextendible null geodesics

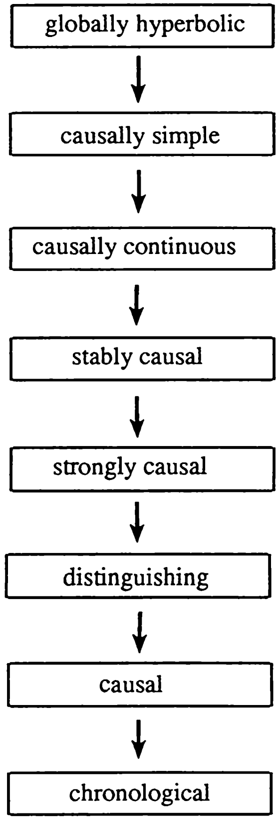


FIGURE 3.3. This diagram illustrates the strengths of the causality conditions used in this book. Global hyperbolicity is the most restrictive causality assumption that we will use.

contained in the set $t = 0$. Recently, the phrase "cluster curve" has been suggested as a more mathematically precise term for the convergence in Definition 3.28.

In contrast, we have the following result for strongly causal space–times.

Lemma 3.29. *Let (M, g) be a strongly causal space–time. If γ is a limit curve of the sequence $\{\gamma_n\}$ of nonspacelike curves, then γ is nonspacelike.*

Proof. Cover γ by a locally finite collection $\{U_k\}$ of convex normal neighborhoods such that for each k, no nonspacelike curve which leaves U_k ever returns. Since the causal relation $\leqslant$ is transitive, it suffices to show that $\gamma \cap U_k$ is nonspacelike for each k.

Let $\{\gamma_m\}$ be a subsequence of $\{\gamma_n\}$ which distinguishes γ. Given any pair of points $p, q \in \gamma \cap U_k$, we may find sequences $\{p_m\}$ and $\{q_m\}$ with $p_m, q_m \in \gamma_m$ for each m and $p_m \to p$, $q_m \to q$. The points p_m and q_m are causally related in U_k for all sufficiently large m by construction of U_k and the assumption that each γ_m is nonspacelike. Taking limits, it follows that p and q are causally related in U_k. Since this holds for each pair $p, q \in \gamma \cap U_k$, it follows that the curve $\gamma \cap U_k$ is nonspacelike. □

The concept of a limit curve is closely related to the Hausdorff closed limit. Let $\{A_n\}$ be an arbitrary sequence of subsets (not necessarily curves) of M. The *Hausdorff upper and lower limits* of $\{A_n\}$ are defined respectively by [cf. Busemann (1955, p. 10)]

$$\limsup\{A_n\} = \{p \in M : \text{ each neighborhood of } p \text{ intersects infinitely many of the sets } A_n\}$$

and

$$\liminf\{A_n\} = \{p \in M : \text{ each neighborhood of } p \text{ intersects all but a finite number of the sets } A_n\}.$$

The Hausdorff upper and lower limits always exist, although they may be empty, and are always closed subsets of M. There is the obvious containment $\liminf\{A_n\} \subseteq \limsup\{A_n\}$. If these two limits are equal, then the *Hausdorff closed limit* of $\{A_n\}$, denoted by $\lim\{A_n\}$, is defined to be $\lim\{A_n\} = \liminf\{A_n\} = \limsup\{A_n\}$.

A limit curve of the sequence of curves $\{\gamma_n\}$ is contained in the Hausdorff upper limit $\limsup\{\gamma_n\}$. Further, a curve γ is a limit curve of the sequence $\{\gamma_n\}$ if and only if $\gamma \subseteq \liminf\{\gamma_m\}$ for some subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$.

We now turn to the proof (Proposition 3.31) of the existence of nonspacelike limit curves for sequences $\{\gamma_n\}$ of nonspacelike curves having points of accumulation. This result, an essential tool of causality theory in general relativity, is a consequence of Arzela's Theorem (Theorem 3.30), which may be invoked since nonspacelike curves satisfy a local Lipschitz condition.

Let U be a convex normal neighborhood of (M, g) with compact closure $\overline{U}$ contained in a chart (V, x) having local coordinates $x = (x_1, \ldots, x_n)$ such that $f = x_1 : U \to \mathbb{R}$ has a timelike gradient ∇f on U. Then f is a time function on U, and whenever c is in the image of f, the level set $f^{-1}(c)$ is a spacelike hypersurface in U. For sufficiently small U there is some constant $K_0 > 0$ such that $g < g_0$ on U where g_0 is the flat metric on U given by

$$g_0 = -K_0 \, dx_1{}^2 + \sum_{j=2}^{n} dx_j{}^2.$$

Furthermore, each nonspacelike curve γ in U joining $p, q \in U$ with $f(p) < f(q)$ can be reparametrized so that γ is given in local coordinates by $\gamma(t) = (t, x_2(t), \ldots, x_n(t))$ for all t with $f(p) \leq t \leq f(q)$. Since γ is nonspacelike for g_0 as well as g, γ satisfies a Lipschitz condition of the form

$$\|\gamma(t_1) - \gamma(t_2)\|_2 \leq K_1 |t_1 - t_2| \tag{3.14}$$

where $K_1 = (K_0 + 1)^{1/2}$. Here, for $p, q \in U$, we use the given local coordinates to define

$$\|p - q\|_2 = \sqrt{\sum_{i=1}^{n} [x_i(p) - x_i(q)]^2},$$

and the constant K_1 depends on g, U, and the choice of local coordinate chart (V, x). This Lipschitz condition implies that γ is differentiable almost everywhere and that $|x_i{}'| \leq K_1$ along γ for all $i = 1, 2, \ldots, n$.

Now let the space–time (M, g) be given an auxiliary complete Riemannian metric h with distance function d_0. By the Hopf–Rinow Theorem, the closed balls $\{q \in M : d_0(p, q) \leq r\}$ are compact for all fixed $p \in M$ and $0 \leq r < \infty$. If the nonspacelike curve $\gamma(t)$ in U is parametrized as $\gamma(t) = (t, x_2(t), \ldots, x_n(t))$ as above, then the length $L_0(\gamma \,|\, [t_1, t_2])$ of γ from t_1 to t_2 with respect to h is given by

$$L_0(\gamma \,|\, [t_1, t_2]) = \int_{t_1}^{t_2} \sqrt{\sum_{i,j} h_{ij} \, x_i{}' \, x_j{}'} \, dt$$

where h_{ij} are the components of h with respect to the local coordinates $x_1, \ldots, x_n$. Since $|x_i'| \leq K_1$, the length $L_0(\gamma \,|\, [t_1, t_2])$ satisfies

$$L_0(\gamma \,|\, [t_1, t_2]) \leq n\, H^{1/2}\, K_1\, |t_1 - t_2| \tag{3.15}$$

where H is the supremum of $|h_{ij}|$ on the compact set $\overline{U}$ for $1 \leq i, j \leq n$. Thus, any nonspacelike curve from the level set $f^{-1}(t_1)$ to the level set $f^{-1}(t_2)$ which lies in U has length bounded by $n\, H^{1/2}\, K_1\, |t_1 - t_2|$. Furthermore, covering (M, g) by a locally finite cover of sets with the properties of U and (V, x) above, it follows that any nonspacelike curve of (M, g) defined on a compact interval of $\mathbb{R}$ must have finite length with respect to h. Thus every nonspacelike curve of (M, g) may be given a parametrization which is an arc length parametrization with respect to h. Also, an inextendible curve γ which has an arc length parametrization with respect to h must be defined on all of $\mathbb{R}$ because d_0 is complete (cf. Lemma 3.65).

We now state a version of Arzela's Theorem which may be established using standard techniques [cf. Munkres (1975, Section 7.5)].

Theorem 3.30. *Let X be a locally compact Hausdorff space with a countable basis, and let (M, h) be a complete Riemannian manifold with distance function d_0. Assume that the sequence $\{f_n\}$ of functions $f_n : X \to M$ is equicontinuous and that for each $x_0 \in X$ the set $\bigcup_n \{f_n(x_0)\}$ is bounded with respect to d_0. Then there exist a continuous function $f : X \to M$ and a subsequence of $\{f_n\}$ which converges to f uniformly on each compact subset of X.*

Using Arzela's Theorem, we may now obtain the next proposition, given in Hawking and Ellis (1973, p. 185), which guarantees the existence of limit curves for a sequence $\{\gamma_n\}$ of nonspacelike curves having points of accumulation.

Proposition 3.31. *Let $\{\gamma_n\}$ be a sequence of (future) inextendible nonspacelike curves in (M, g). If p is an accumulation point of the sequence $\{\gamma_n\}$, then there is a nonspacelike limit curve γ of the sequence $\{\gamma_n\}$ such that $p \in \gamma$ and γ is (future) inextendible.*

Proof. We will give the proof only for inextendible curves since the proof for future inextendible curves is similar.

Let h be an auxiliary complete Riemannian metric for M with distance function d_0 as above, and give each γ_n an arc length parametrization with respect to h. Then the domain of each γ_n is $\mathbb{R}$ as each curve is assumed to be inextendible. Shifting parametrizations if necessary, we may then choose a subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$ such that $\gamma_m(0) \to p$ as $m \to \infty$ since p is an accumulation point of the sequence γ_n. Using the fact that each γ_m has an arc length parametrization with respect to h, we obtain

$$d_0(\gamma_m(t_1), \gamma_m(t_2)) \leq |t_1 - t_2| \tag{3.16}$$

for each m and $t_1, t_2 \in \mathbb{R}$. Thus the curves $\{\gamma_m\}$ form an equicontinuous family. Furthermore, since $\gamma_m(0) \to p$, there exists an N such that $d_0(\gamma_m(0), p) < 1$ whenever $m \geq N$. This implies that for each fixed $t_0 \in \mathbb{R}$, the curve $\gamma_m \mid [-t_0, t_0]$ of the subsequence lies in the compact set $\{q \in M : d_0(p, q) \leq t_0 + 1\}$ whenever $m \geq N$. Hence the family $\{\gamma_m\}$ satisfies the hypotheses of Theorem 3.30, and we thus obtain a (continuous) curve $\gamma : \mathbb{R} \to M$ and a subsequence $\{\gamma_k\}$ of the subsequence $\{\gamma_m\}$ such that $\{\gamma_k\}$ converges to γ uniformly on each compact subset of $\mathbb{R}$. Clearly, $\gamma_k(0) \to p = \gamma(0)$. The convergence of $\{\gamma_k\}$ to γ also yields the inequality $d_0(\gamma(t_1), \gamma(t_2)) \leq |t_1 - t_2|$ for all $t_1, t_2 \in \mathbb{R}$. It remains to show that γ is nonspacelike and inextendible.

To show that γ is nonspacelike, fix $t_1 \in \mathbb{R}$ and let U be a convex normal neighborhood of (M, g) containing $\gamma(t_1)$. Choose $\delta > 0$ such that the set $\{q \in M : d_0(\gamma(t_1), q) < \delta\}$ is contained in U. If $t_1 < t_2 < t_1 + \delta$, then (3.16) and the uniform convergence on compact subsets yields that for all large k, the set $\gamma_k[t_1, t_2]$ lies in U. Using $\gamma_k(t_1) \to \gamma(t_1)$, $\gamma_k(t_2) \to \gamma(t_2)$, $\gamma_k(t_1) \leqslant_U \gamma_k(t_2)$ for all large k, and the fact that U is a convex normal neighborhood, we obtain that $\gamma(t_1) \leqslant_U \gamma(t_2)$. Thus $\gamma \mid [t_1, t_2]$ is a future directed nonspacelike curve in U [cf. Hawking and Ellis (1973, Proposition 4.5.1)]. It follows that γ is a future directed nonspacelike curve in (M, g).

It remains to show that γ is inextendible. We will give the proof only of the future inextendibility since the past inextendibility may be proven similarly. To this end, assume that γ is not future inextendible. Then $\gamma(t) \to q_0 \in M$ as $t \to \infty$. Let U' be a convex normal neighborhood of q_0 such that $\overline{U'}$ is a compact set contained in a chart (V, x) of M with local coordinates $(x_1, \ldots, x_n)$

such that $f = x_1 : U' \to \mathbb{R}$ is a time function for U'. An inequality of the form of (3.15) shows that if $\gamma | [t_1, \infty) \subseteq U'$, then no nonspacelike curve in U' from the level set $f^{-1}(f(\gamma(t_1)))$ to the level set $f^{-1}(f(q_0))$ can have arc length with respect to h greater than some number $\delta' > 0$. On the other hand, for sufficiently large k we must have $\gamma_k[t_1+1, t_1+\delta'+2] \subseteq f^{-1}([f(\gamma(t_1)), f(q_0)])$. Since the length $L_0(\gamma_k[t_1 + 1, t_1 + \delta' + 2]) = \delta' + 1$ for all k, this yields a contradiction. □

Even if all of the inextendible nonspacelike curves of the sequence $\{\gamma_n\}$ are parametrized by arc length with respect to a complete Riemannian metric h, the limit curve γ obtained in the proof of Proposition 3.31 need not be parametrized by arc length. This is a consequence of the fact that the Riemannian length functional, while lower semicontinuous, is not upper semicontinuous in the topology of uniform convergence on compact subsets. Even though the curve γ constructed in the proof of Proposition 3.31 need not be parametrized by arc length, the curve γ will still be defined on all of $\mathbb{R}$ provided each γ_n is inextendible. Furthermore, if (M, g) is strongly causal, the Hopf–Rinow Theorem and Proposition 3.13 imply that $d_0(\gamma(0), \gamma(t)) \to \infty$ as $|t| \to \infty$. Here, d_0 denotes the complete Riemannian distance function induced on M by h as above. An alternative treatment of the technicalities of Proposition 3.31, closer in spirit to that given in Hawking and Ellis, may be found in O'Neill (1983, p. 404) in the section on "quasi-limits."

In the globally hyperbolic case, a stronger version of Proposition 3.31 may be obtained.

Corollary 3.32. *Let (M, g) be globally hyperbolic. Suppose that $\{p_n\}$ and $\{q_n\}$ are sequences in M converging to p and q in M respectively, with $p \leqslant q$, $p \neq q$, and $p_n \leqslant q_n$ for each n. Let γ_n be a future directed nonspacelike curve from p_n to q_n for each n. Then there exists a future directed nonspacelike limit curve γ of the sequence $\{\gamma_n\}$ which joins p to q.*

Proof. Let h be an auxiliary complete Riemannian metric on M with length functional L_0. Choose a finite cover of the compact set $J^+(p) \cap J^-(q)$ by convex normal neighborhoods $U_1, U_2, \ldots, U_k$, each of which has compact closure and such that no nonspacelike curve which leaves any U_i ever returns to that

U_i. As in the proof of Proposition 3.31, there exists a number N_i for each i such that each nonspacelike curve $\gamma : [a, b] \to U_i$ has length less than N_i with respect to h [cf. equation (3.15)]. Thus if $U = U_1 \cup \cdots \cup U_k$ and $N = N_1 + \cdots + N_k$, every nonspacelike curve $\gamma : [a, b] \to U$ must satisfy $L_0(\gamma) \leq N$.

Extend each given nonspacelike curve γ_n to a future inextendible nonspacelike curve, also denoted by γ_n. We may assume that each $\gamma_n : [0, \infty) \to M$ has been parametrized by arc length with respect to h [cf. equation (3.15)]. Thus if $U = U_1 \cup \cdots \cup U_k$ and Proposition 3.31 is applied to $\{\gamma_n\}$ with accumulation point p of $\{\gamma_n(0) = p_n\}$, then there exist a future inextendible nonspacelike limit curve $\gamma : [0, \infty) \to M$ with $\gamma(0) = p$ and a subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$ such that $\gamma_m \to \gamma$ uniformly on compact subsets of $[0, \infty)$. Using $\gamma_m(t_m) = q_m$ for $0 < t_m \leq N$ and $q_m \to q$, we conclude that γ passes through q for some parameter value τ which satisfies $0 < \tau \leq N$. It follows that $\gamma | [0, \tau]$ is a nonspacelike limit curve of $\{\gamma_m | [0, t_m]\}$ which joins p to q. □

We now consider convergence in the C^0 topology [cf. Penrose (1972, p. 49)].

Definition 3.33. (*Convergence of Curves in C^0 Topology*) Let γ and all curves of the sequence $\{\gamma_n\}$ be defined on the closed interval $[a, b]$. The sequence $\{\gamma_n\}$ is said to *converge to γ in the C^0 topology on curves* if $\gamma_n(a) \to \gamma(a)$, $\gamma_n(b) \to \gamma(b)$, and given any open set V containing γ, there is an integer N such that $\gamma_n \subseteq V$ for all $n \geq N$.

Any space–time contains a sequence $\{\gamma_n\}$ that has a limit curve γ, yet $\{\gamma_n\}$ does not converge to γ in the C^0 topology. For, let $\alpha, \beta : [0, 1] \to M$ be any two future directed timelike curves with $\alpha([0, 1]) \cap \beta([0, 1]) = \emptyset$. Set

$$\gamma_n = \begin{cases} \alpha & \text{if} \quad n = 2m, \\ \beta & \text{if} \quad n = 2m - 1. \end{cases}$$

Then $\{\gamma_n\}$ does not converge to either α or β in the C^0 topology. However, the subsequence $\{\gamma_{2n}\}$ (respectively, $\{\gamma_{2n-1}\}$) of $\{\gamma_n\}$ converges to α (respectively, β) in the C^0 topology. A space–time which is not strongly causal may also contain a sequence $\{\gamma_n\}$ of nonspacelike curves which has a nonspacelike limit curve γ, yet no subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$ converges to γ in the C^0 topology on

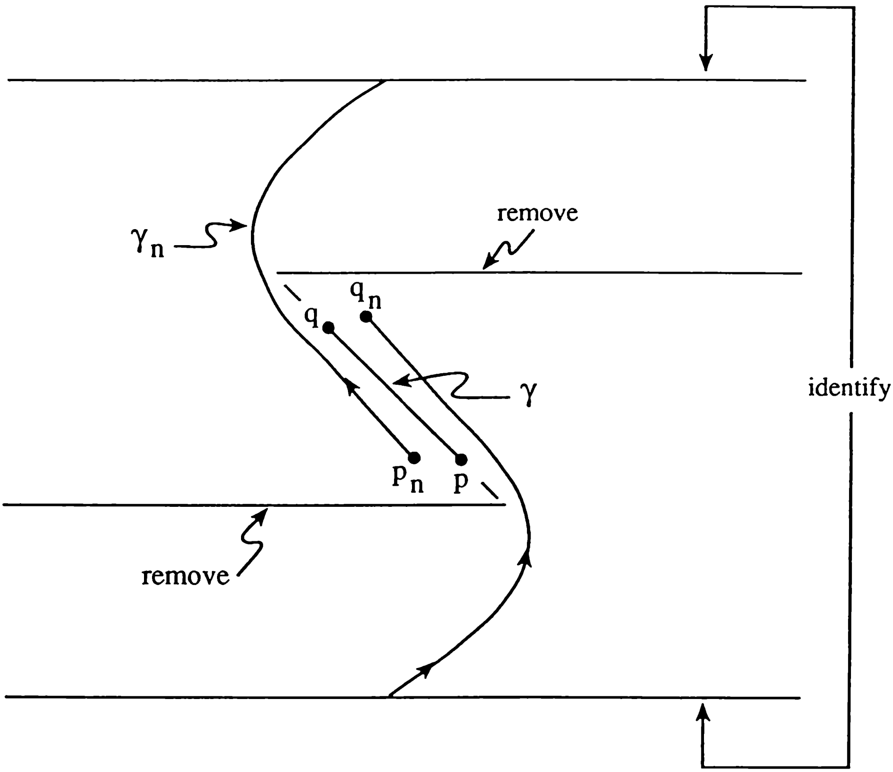


FIGURE 3.4. A causal space–time (M, g), in which a sequence $\{\gamma_n\}$ of nonspacelike curves may have a limit curve γ yet fail to have a subsequence which converges to γ in the C^0 topology on curves, may be formed from a subset of Minkowski space as shown.

curves. This is illustrated in Figure 3.4 [cf. Hawking and Ellis (1973, p. 193) for a discussion of the causal properties of this example].

Conversely, a sequence of nonspacelike curves $\{\gamma_n\}$ may converge in the C^0 topology to some nonspacelike curve γ but fail to have γ as a limit curve. This may be seen on the cylinder $M = S^1 \times \mathbb{R}$ with the Lorentzian metric $ds^2 = d\theta\, dt$. Let γ_n be the segment on the generator $\theta = 0$ given by $\gamma_n(t) = (0, t)$ for $0 \leq t \leq 1$ and for all n. If γ is the piecewise smooth nonspacelike curve obtained by going around the circle on the null geodesic $t = 0$ and then up the generator $\theta = 0$ from $t = 0$ to $t = 1$, then $\{\gamma_n\}$ converges to γ in the C^0 topology, but γ is not a limit curve of $\{\gamma_n\}$ (cf. Figure 3.5).

In strongly causal space–times, however, these two types of convergence are almost equivalent for sequences of nonspacelike curves [cf. Beem and Ehrlich (1979a, p. 164)]. A more precise statement is given by the following result.

Proposition 3.34. *Let (M, g) be a strongly causal space–time. Suppose that $\{\gamma_n\}$ is a sequence of nonspacelike curves defined on $[a, b]$ such that $\gamma_n(a) \to p$ and $\gamma_n(b) \to q$. A nonspacelike curve $\gamma : [a, b] \to M$ with $\gamma(a) = p$ and $\gamma(b) = q$ is a limit curve of $\{\gamma_n\}$ iff there is a subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$ which converges to γ in the C^0 topology on curves.*

Proof. ($\Rightarrow$) We may assume without loss of generality that γ and $\{\gamma_n\}$ are all future directed curves. Let V be any open set with $\gamma \subseteq V$. Cover the compact image of γ with convex normal neighborhoods $W_1, W_2, \ldots, W_k$ such that each $W_i \subseteq V$ and no nonspacelike curve which leaves W_i ever returns to W_i. There exists a subdivision $a = t_0 < t_1 < \cdots < t_j = b$ of $[a, b]$ such that for all $0 \leq i \leq j - 1$, each pair $\gamma(t_i)$, $\gamma(t_{i+1})$ lies in some W_h. Here $h = h(i)$ and $1 \leq h(i) \leq k$ for all i. Let $\{\gamma_m\}$ be a subsequence that distinguishes γ as a limit curve. For each m, let $p(0, m) = \gamma_m(a)$ and $p(j, m) = \gamma_m(b)$. Furthermore, for each fixed i with $0 < i < j$, choose $p(i, m) \in \gamma_m$ such that $\{p(i, m)\}$ converges to $\gamma(t_i)$. Since $\gamma(t_{i+1})$ lies in the causal future of $\gamma(t_i)$ and M is strongly causal, the point $p(i + 1, m)$ lies in the causal future of $p(i, m)$ for all m larger than some N_1. Also, there is some N_2 such that $p(i, m)$ and $p(i+1, m)$ lie in $W_{h(i)}$ for all $0 \leq i \leq j-1$ and $m \geq N_2$. Let $N = \max\{N_1, N_2\}$. The portion of γ_m joining $p(i, m)$ to $p(i + 1, m)$ must lie entirely in $W_{h(i)}$ for $m \geq N$ because no nonspacelike curve can leave W_h and return. It follows that $\gamma_m \subseteq W_1 \cup \cdots \cup W_k \subseteq V$ for all $m \geq N$ as required.

($\Leftarrow$) Let $\{\gamma_m\}$ be a subsequence of $\{\gamma_n\}$ converging to γ in the C^0 topology on curves. Define $A = \{t_0 \in [a, b] :$ each point of $\gamma | [a, t_0]$ is a limit point of the given subsequence $\}$. We wish to show that $A = [a, b]$. Clearly, $\gamma_m(a) \to \gamma(a)$ implies $a \in A$. If $\tau = \sup\{t_0 : t_0 \in A\}$, then for each $a \leq t < \tau$ the point $\gamma(t)$ is a limit point of the subsequence $\{\gamma_m\}$. To show $\tau \in A$ we assume

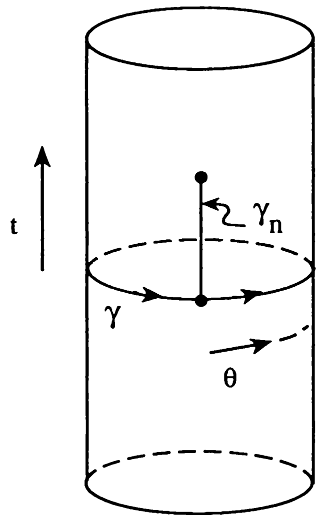


FIGURE 3.5. In chronological space–times, a sequence of nonspacelike curves $\{\gamma_n\}$ may converge to the nonspacelike curve γ in the C^0 topology on curves, and yet γ may fail to be a limit curve of any subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$. The curves γ_n are segments on the line $\theta = 0$ from $t = 0$ to $t = 1$. The curve γ goes around the cylinder once, then traverses γ_n.

$\tau > a$ and let $\{t_k\}$ be a sequence with $t_k \to \tau^-$. Each neighborhood $U(\gamma(\tau))$ of $\gamma(\tau)$ is also a neighborhood of $\gamma(t_k)$ for sufficiently large k and hence must intersect all but a finite number of curves of the subsequence $\{\gamma_m\}$. Thus $\gamma(\tau)$ is a limit point of $\{\gamma_m\}$, and A must be a closed subinterval of $[a, b]$. Assume that $\tau < b$. Using the strong causality of (M, g), we may find a convex normal neighborhood V of $\gamma(\tau)$ such that no nonspacelike curve of (M, g) which leaves V ever returns. Letting V be sufficiently small we may assume that $(V, g|_V)$ is globally hyperbolic and that $f : V \to \mathbb{R}$ is a Cauchy time function for $(V, g|_V)$ with $f(V) = \mathbb{R}$ and $f(\gamma(\tau)) = 0$. We may also assume $\gamma(b) \notin V$. Then each inextendible nonspacelike curve of (M, g) which has a nonempty intersection with V must intersect each Cauchy surface $f^{-1}(s)$ exactly once. Fix s with $0 < s < \infty$, and define $x(s) = \gamma \cap f^{-1}(s)$. This intersection exists because $\gamma(\tau) \in V$ and $\gamma(b) \notin V$. Since $\gamma(\tau)$ and $\gamma(b)$ are limit points of $\{\gamma_m\}$, the curves γ_m must have a nonempty intersection with $f^{-1}(s)$ for all sufficiently large m. Set $x_m(s) = \gamma_m \cap f^{-1}(s)$ for all such m. In order to verify that $x_m(s) \to x(s)$, we observe that for each neighborhood W of γ the points $x_m(s)$ must lie in $W \cap f^{-1}(s)$ for all large m (cf. Figure 3.6). This shows that each $x(s)$ is a limit point of the subsequence $\{\gamma_m\}$. Consequently, the set A contains numbers greater than τ, in contradiction to the definition of τ. We conclude $A = [a, b]$ which shows γ is a limit curve of the subsequence $\{\gamma_m\}$. □

Let $\gamma : [a, b] \to (M, g)$ be a nonspacelike curve in a strongly causal space–time (M, g). Choose a compact subset K of M such that $\gamma \subseteq \mathrm{Int}(K)$. Let the nonspacelike curves which are contained in K be given the C^0 topology. It is known [cf. Penrose (1972, p. 54)] that the Lorentzian arc length functional $L(\gamma)$ [cf. Chapter 4, equation (4.1)] is upper semicontinuous with respect to the C^0 topology on curves [cf. Busemann (1967, p. 10)]. This is the analogue of the well-known result that the Riemannian arc length functional is lower semicontinuous.

Remark 3.35. Let (M, g) be strongly causal, and let γ be a given non-spacelike curve in (M, g). If the sequence $\{\gamma_n\}$ of nonspacelike curves converges to γ in the C^0 topology on curves, then

$$L(\gamma) \geq \limsup L(\gamma_n).$$

3.4 Two-Dimensional Space–times

In this section we consider the topological and causal structures of two-dimensional Lorentzian manifolds. Using the pair of null vector fields generated by the tangent vectors to the two null geodesics passing through each point of M, we show that the universal covering manifold of any two-dimensional Lorentzian manifold is homeomorphic to $\mathbb{R}^2$. We then show that any two-dimensional Lorentzian manifold homeomorphic to $\mathbb{R}^2$ is stably causal. In

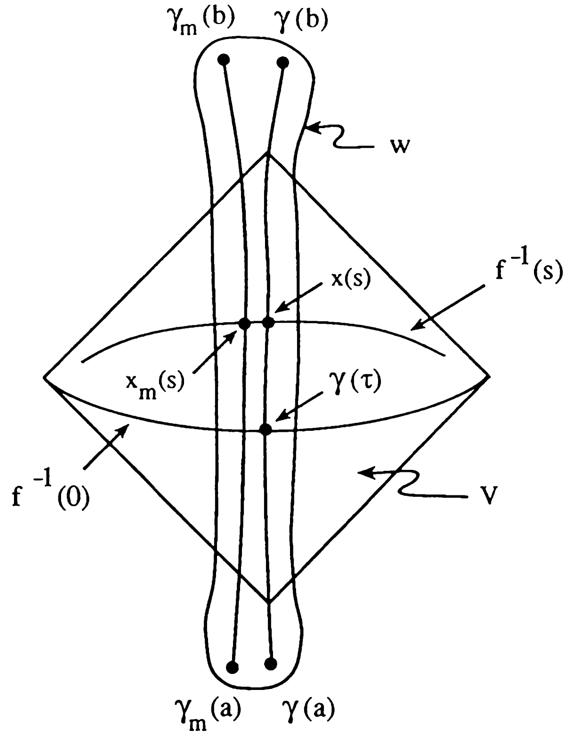


FIGURE 3.6. In the proof of Proposition 3.34 the globally hyperbolic neighborhood V of $\gamma(\tau)$ has a Cauchy time function $f : V \to \mathbb{R}$ with $f(\gamma(\tau)) = 0$. For all large m the curves γ_m must intersect the Cauchy surface $f^{-1}(s)$ at a single point $x_m(s)$. If W is any neighborhood of γ, then $x_m(s) \in W \cap f^{-1}(s)$ for all large m. We may choose W such that $W \cap f^{-1}(s)$ is as small a neighborhood of $x(s) = \gamma \cap f^{-1}(s)$ in $f^{-1}(s)$ as we wish. Thus $x_m(s) \to x(s)$, and $x(s)$ must be a limit curve of the subsequence $\{\gamma_m\}$.

particular, every simply connected two-dimensional Lorentzian manifold is causal. Thus *no* Lorentzian metric for $\mathbb{R}^2$ has any closed nonspacelike curves. It should also be noted that two- (but not higher) dimensional Lorentzian manifolds have the property that $(M, -g)$ is also Lorentzian. This is sometimes useful in obtaining results about all geodesics in (M, g) from results valid in higher dimensions only for nonspacelike geodesics.

Let (M, g) be an arbitrary two-dimensional Lorentzian manifold, and fix a point $p \in M$. Choose a convex normal neighborhood $U(p)$ based at p, and consider the following method of assigning local coordinates to points in $U(p)$ sufficiently close to p. Let the two null geodesics γ_1 and γ_2 through p be given parametrizations $\gamma_1 : (-\epsilon_1, \epsilon_1) \to U(p)$ and $\gamma_2 : (-\epsilon_2, \epsilon_2) \to U(p)$ with $\gamma_1(0) = \gamma_2(0) = p$. For each point $q \in U(p)$ sufficiently close to p, the two null geodesics through q will intersect γ_1 and γ_2 in $U(p)$ at unique points $\gamma_1(t_0)$ and $\gamma_2(s_0)$ respectively. Assign coordinates (t_0, s_0) to q. In these coordinates the null geodesics near p are contained in sets of the form $t = t_0$ or $s = s_0$. We have established

Lemma 3.36. *Let (M, g) be a two-dimensional Lorentzian manifold. Then each $p \in M$ has local coordinates $x = (x_1, x_2)$ with $x(p) = 0$ such that each null geodesic in this neighborhood is contained in a set of the form $x_1 =$ constant or $x_2 =$ constant.*

Suppose X is a future directed timelike vector field on M. Then at each $p \in M$, there are two uniquely defined future directed null vectors $n_1, n_2 \in T_pM$ such that $X(p) = n_1 + n_2$. Clearly, a sufficiently small neighborhood $U(p)$ of p may be found such that n_1 and n_2 may be extended to continuous null vector fields X_1, X_2 defined on $U(p)$ with $X(q) = X_1(q) + X_2(q)$ for all $q \in U(p)$. If M is simply connected, we now show X_1 and X_2 can be extended to all of M.

Proposition 3.37. *Let (M, g) be a simply connected Lorentzian manifold of dimension two. Then two smooth nonvanishing null vector fields X_1 and X_2 may be defined on M such that X_1 and X_2 are linearly independent at each point of M.*

Proof. Since M is simply connected, (M, g) is time orientable. Thus we may choose a smooth future directed timelike vector field X on M.

Fix a base point $p_0 \in M$, and let $X(p_0) = n_1 + n_2$ as above. Given any other point $q \in M$, let $\gamma : [0, 1] \to M$ be a curve from p_0 to q. There is exactly one way to define continuous null vector fields X_1 and X_2 along γ such that

$X_1(0) = n_1$, $X_2(0) = n_2$, and $X(\gamma(t)) = X_1(t) + X_2(t)$ for all $t \in [0,1]$. If $\eta : [0,1] \to M$ is any other curve from p_0 to q, then γ and η are homotopic since M is assumed to be simply connected. Hence if Y_1 and Y_2 were null vector fields along η with $Y_1(0) = n_1$ and $Y_2(0) = n_2$, we would have $Y_1(1) = X_1(1)$ and $Y_2(1) = X_2(1)$ by standard homotopy arguments. Thus this construction produces a pair of continuous vector fields X_1 and X_2 on M which are linearly independent at each point. □

Corollary 3.38. *Let (M, g) be any two-dimensional Lorentzian manifold. Then the universal Lorentzian covering manifold $(\widetilde{M}, \widetilde{g})$ of (M, g) is homeomorphic to $\mathbb{R}^2$.*

Proof. Since $\widetilde{M}$ is simply connected and two-dimensional, $\widetilde{M}$ is homeomorphic to $\mathbb{R}^2$ or S^2. But since the Euler characteristic of S^2 is nonzero, S^2 does not admit any nowhere zero continuous vector fields. □

Recall that an *integral curve* for a smooth vector field X on M is a smooth curve γ such that $\gamma'(t) = X(\gamma(t))$ for all t in the domain of γ [cf. Kobayashi and Nomizu (1963, p. 12)]. The following result is well known [cf. Hartman (1964, p. 156)].

Proposition 3.39. *Let X be a smooth nonvanishing vector field on $\mathbb{R}^2$, and let $\gamma : (a, b) \to \mathbb{R}^2$ be a maximal integral curve of X. Then $\gamma(t)$ does not remain in any compact subset of $\mathbb{R}^2$ as $t \to a^+$ (or $t \to b^-$).*

Now assume that (M, g) is a Lorentzian manifold homeomorphic to $\mathbb{R}^2$. Let X_1, X_2 be the null vector fields on M given by Proposition 3.37. Clearly, each null geodesic of (M, g) may be reparametrized to an integral curve of X_1 or X_2. Equivalently, the integral curves of X_1 and X_2 are said to be *null pregeodesics*. Suppose $\gamma : (a, b) \to M$ is an inextendible null geodesic which may be reparametrized to an integral curve of X_1. If $\gamma(t_1) = \gamma(t_2)$ for some $t_1 \neq t_2$, then since both $\gamma'(t_1)$ and $\gamma'(t_2)$ are scalar multiples of $X_1(\gamma(t_1))$, it follows from geodesic uniqueness that γ is a smooth closed geodesic. However, this is impossible by Proposition 3.39. Thus Proposition 3.39 has the following corollary.

Corollary 3.40. *If (M, g) is a Lorentzian manifold homeomorphic to $\mathbb{R}^2$, then (M, g) contains no closed null geodesics. Moreover, every inextendible null geodesic $\gamma : (a, b) \to M$ is injective and hence contains no loops.*

A family F of inextendible null geodesics is said to cover a manifold M *simply* if each point $p \in M$ lies on exactly one null geodesic of F. Suppose (M, g) is a Lorentzian manifold homeomorphic to $\mathbb{R}^2$. Then the integral curves of the null vector field X_1 (respectively, X_2) given in Proposition 3.37 may be reparametrized to define a family F_1 (respectively, F_2) of geodesics on M. Each family F_i covers M since $X_i(p) \neq 0$ for $i = 1, 2$ and all $p \in M$. Furthermore, since exactly one integral curve of X_i passes through any $p \in M$, each family F_i covers M simply. Consequently, Proposition 3.37 implies [cf. Beem and Woo (1969, p. 51)]

Proposition 3.41. *Let (M, g) be a Lorentzian manifold homeomorphic to $\mathbb{R}^2$. Then the inextendible null geodesics of (M, g) may be partitioned into two families F_1 and F_2 such that each of these families covers M simply.*

Let $\gamma : (a, b) \to M$ be an inextendible timelike curve, and let $c : (\alpha, \beta) \to M$ be an inextendible null geodesic. Obviously, in arbitrary two-dimensional Lorentzian manifolds, γ and c may intersect more than once. However, if M is homeomorphic to $\mathbb{R}^2$, γ and c intersect in at most one point [cf. Beem and Woo (1969, p. 52), Smith (1960b)].

Proposition 3.42. *Let (M, g) be a Lorentzian manifold homeomorphic to $\mathbb{R}^2$. Then each timelike curve intersects a given null geodesic at most once.*

Proof. Let c_0 be an inextendible future directed null geodesic in M, which we may assume belongs to the family F_1 defined by the null vector field X_1 as above. Suppose that σ is a future directed timelike curve in M which intersects c_0 twice (possibly at the same point). We may then find $a, b \in \mathbb{R}$ with $a < b$ such that $\sigma(a)$, $\sigma(b)$ lie on c_0 and $\sigma(t) \notin c_0$ for $a < t < b$. Since σ is timelike, σ is locally one-to-one. Hence if $\sigma \mid [a, b]$ is not one-to-one, σ contains at worst closed timelike loops. Using one of these loops, it is possible to find α, $\beta \in \mathbb{R}$ with $a < \alpha < \beta < b$ and a second null geodesic $c_1 \in F_1$ such that $\sigma \mid [\alpha, \beta]$ is one-to-one, $\sigma(\alpha)$ and $\sigma(\beta)$ lie on c_1, and $\sigma(t) \notin c_1$ for $\alpha < t < \beta$.

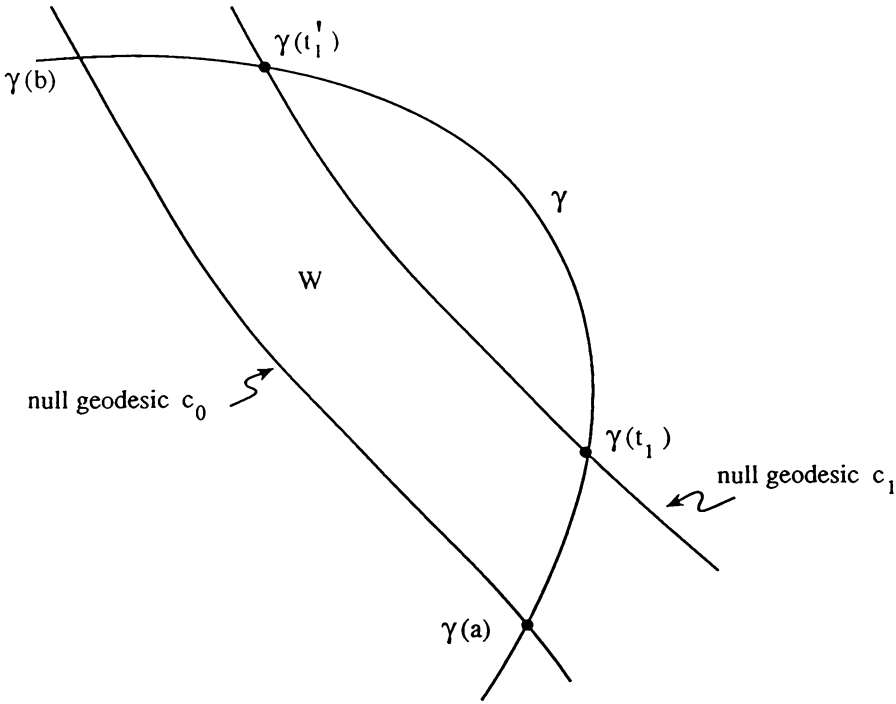


FIGURE 3.7. In a Lorentzian manifold homeomorphic to $\mathbb{R}^2$, the timelike curve γ is assumed to cross the null geodesic c_0 at $\gamma(a)$ and $\gamma(b)$. The null geodesic c_1 enters W at $\gamma(t_1)$ and first leaves at $\gamma(t_1')$ where $t_1' > t_1$.

We will show below that if $\gamma : [a,b] \to M$ is an injective future directed timelike curve, there is no $c \in F_1$ such that $\gamma(a)$ and $\gamma(b)$ lie on c but $\gamma(t) \notin c$ for $a < t < b$. If the original timelike curve $\sigma \mid [a,b]$ is injective, this argument applied to $\sigma \mid [a,b]$ yields the desired contradiction. If $\sigma \mid [a,b]$ is not injective but intersects c_0 at $\sigma(a)$ and $\sigma(b)$, then this argument applied to c_1 and $\sigma \mid [\alpha,\beta]$ yields the desired contradiction.

Thus the theorem will be established if we show that it is impossible to find an injective future directed timelike curve $\gamma : [a,b] \to M$ with $\gamma(a)$ and $\gamma(b)$ on c_0 and $\gamma(t) \notin c_0$ for $a < t < b$. Traversing γ from $\gamma(a)$ to $\gamma(b)$ and then the portion of c_0 from $\gamma(b)$ to $\gamma(a)$ yields a closed Jordan curve which encloses a set W with $\overline{W}$ compact (cf. Figure 3.7). Let U be a convex normal neighborhood based on $\gamma(a)$. Choose t_1 with $a < t_1 < b$ and $\gamma(t_1) \in U$. Let c_1 be the null geodesic in F_1 passing through $\gamma(t_1)$. Since c_1 may be reparametrized to be an integral curve of X_1 and c_1 enters $\overline{W}$ at $\gamma(t_1)$, it follows by Proposition 3.39 that c_1 leaves W at some point $\gamma(t_1')$ with $t_1' > t_1$. As c_0 intersects γ at $\gamma(b)$, we must have $t_1' < b$. In particular, $[t_1, t_1'] \subseteq (a,b)$. Hence we have found a closed interval $[t_1, t_1'] \subseteq (a,b)$ such that $\gamma([t_1, t_1']) \subseteq \overline{W}$ and γ intersects the null geodesic $c_1 \in F_1$ at $\gamma(t_1)$ and $\gamma(t_1')$ (cf. Figure 3.7).

We may now form a second closed Jordan curve by traversing γ from t_1 to t_1' followed by the portion of c_1 from $\gamma(t_1')$ to $\gamma(t_1)$. Repeating the argument of the preceding paragraph, we obtain a closed interval $[t_2, t_2'] \subseteq (t_1, t_1')$ such that the timelike curve $\gamma | [t_2, t_2']$ intersects a null geodesic c_2 in the family F_1 at $\gamma(t_2)$ and $\gamma(t_2')$ and such that $\gamma(t_2)$ is contained in a convex normal neighborhood of $\gamma(t_1)$. Inductively, we can construct a nested sequence of intervals $[t_{k+1}, t_{k+1}'] \subseteq (t_k, t_k')$ such that $\gamma(t_{k+1})$ lies in a convex normal neighborhood of $\gamma(t_k)$ and $\gamma | [t_{k+1}, t_{k+1}']$ intersects a null geodesic $c_{k+1} \in F_1$ at $\gamma(t_{k+1})$ and $\gamma(t_{k+1}')$. Moreover, the intervals $[t_k, t_k']$ may be chosen such that $\bigcap_{k=1}^{\infty} [t_k, t_k'] = \{t_0\}$ for some $t_0 \in (a,b)$. We thus have constructed two sequences $t_k \uparrow t_0$ and $t_k' \downarrow t_0$ such that the timelike curve γ intersects a null geodesic in F_1 at both $\gamma(t_k)$ and $\gamma(t_k')$ for each $k \geq 1$. But this is impossible by Proposition 3.4. Hence the geodesic $\gamma : [a,b] \to M$ intersects c_0 at most once. □

Theorem 3.43. *Let (M,g) be a Lorentzian manifold homeomorphic to $\mathbb{R}^2$. Then (M,g) is stably causal.*

Proof. Recall that $g \in \text{Lor}(M)$ is stably causal if there is a fine C^0 neighborhood U of g in $\text{Lor}(M)$ such that all metrics in U are causal. Since strongly causal implies causal, it will thus follow that all metrics in $\text{Lor}(M)$ are stably causal if all metrics in $\text{Lor}(M)$ are strongly causal. Thus to prove the theorem, it is enough to show that if g is any Lorentzian metric for M, then (M, g) is strongly causal.

Thus suppose that g is a Lorentzian metric for M such that (M, g) is not strongly causal. Then there is some $p \in M$ such that strong causality fails at p. Let (U, x) be a chart about p, guaranteed by Lemma 3.36, such that the null geodesics in U lie on the sets $x_1 =$ constant and $x_2 =$ constant. Since strong causality fails at p, there are arbitrarily small neighborhoods V of p with $V \subseteq U$ and timelike curves which begin at p, leave V, and then return to V. By Proposition 3.42, there are no closed timelike curves through p. Thus if γ is a future directed timelike curve with $\gamma(0) = p$ which leaves V and then returns, we have $\gamma(t) \neq p$ for all $t > 0$. Since the null geodesics in U through p are given by $x_1 = 0$ and by $x_2 = 0$ in the local coordinates $x = (x_1, x_2)$ for U, it follows that γ may be deformed to intersect one of the null geodesics through p upon returning to V (cf. Figure 3.8). Hence γ intersects a null geodesic in F_1 or F_2 twice, contradicting Proposition 3.42. □

Corollary 3.44. *No Lorentzian metric for $\mathbb{R}^2$ contains any closed nonspacelike curves.*

A different proof of the result that any simply connected Lorentzian two-manifold is strongly causal may be found in O'Neill (1983). For $n \geq 3$, Lorentzian metrics which are not chronological and hence not strongly causal may be constructed on $\mathbb{R}^n$.

For every two-dimensional Lorentzian manifold (M, g), there is an associated Lorentzian manifold $(M, -g)$. The timelike curves of $(M, -g)$ are the spacelike curves of (M, g) and vice versa. Using $M = \mathbb{R}^2$ and applying Corollary 3.44 to $(M, -g)$ we obtain

Corollary 3.45. *No Lorentzian metric for $\mathbb{R}^2$ contains any closed spacelike curves.*

If (M, g) is two-dimensional and both (M, g) and $(M, -g)$ are stably causal, then using techniques given in Beem (1976a), one may show there is some smooth conformal factor $\Omega : M \to (0, \infty)$ such that the manifold $(M, \Omega g)$ is geodesically complete. This yields the following corollary.

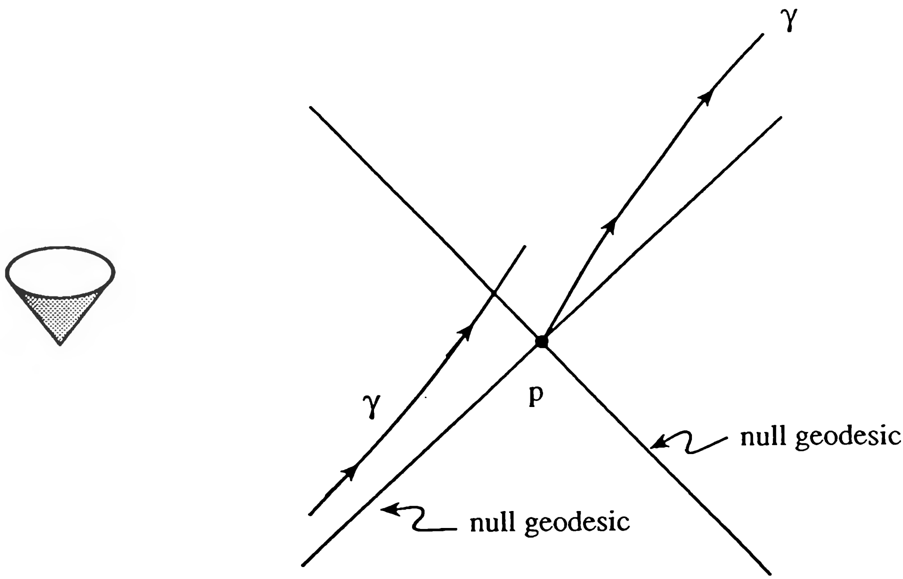


FIGURE 3.8. (M, g) is a two-dimensional space–time such that strong causality fails at p. There is a future directed timelike curve γ which starts at p, later returns close to p, and crosses one of the null geodesics through p.

Corollary 3.46. *Let (M, g) be a Lorentzian manifold homeomorphic to $\mathbb{R}^2$. Then there is a smooth conformal factor $\Omega : M \to (0, \infty)$ such that $(M, \Omega g)$ is geodesically complete.*

There are examples of two-dimensional space–times such that no global conformal change makes them nonspacelike geodesically complete (cf. Section 6.2). Thus, Corollary 3.46 cannot be extended to all two-dimensional space–times by covering space arguments. The recent monograph by Weinstein (1993) contains many further interesting results on Lorentzian surfaces inspired in part by Kulkarni's (1985) study of the conformal boundary for such a surface [cf. Smyth and Weinstein (1994)].

3.5 The Second Fundamental Form

Let N be a smooth submanifold of the Lorentzian manifold (M, g). If $i : N \to M$ denotes the inclusion map, then we may regard T_pN as being a subspace of T_pM by identifying $i_{*_p}(T_pN)$ with T_pN. Let $g_0 = i^*g$ denote the pullback of the Lorentzian metric g for M to a symmetric tensor field on N. Under the identification of T_pN and $i_{*_p}(T_pN)$, we may also identify g_0 at p and $g \mid T_pN \times T_pN$ for all $p \in N$. This identification will be used throughout this section.

Definition 3.47. (*Nondegenerate Submanifold*) The submanifold N of (M, g) is said to be *nondegenerate* if for each $p \in N$ and nonzero $v \in T_pN$, there exists some $w \in T_pN$ with $g(v, w) \neq 0$. If, in addition, $g \mid T_pN \times T_pN$ is positive definite for each $p \in N$, then N is said to be a *spacelike submanifold.* If $g \mid T_pN \times T_pN$ is a Lorentzian metric for each $p \in N$, then N is said to be a *timelike submanifold.*

For the rest of this section, we will suppose that N is a nondegenerate submanifold. Thus for each $p \in N$, there is a well-defined subspace $T_p^{\perp}N$ of T_pM given by

$$T_p^{\perp}N = \{v \in T_pM : g(v, w) = 0 \text{ for all } w \in T_pN\}$$

which has the property that $T_p^{\perp}N \cap T_pN = \{0\}$. Consequently, there is a well-defined orthogonal projection map $P : T_pM \to T_pN$. The connection ∇ on (M, g) may be projected to a connection ∇^0 on N by defining $\nabla^0_X Y = P(\nabla_X Y)$ for vector fields X, Y tangent to N. It is easily verified that ∇^0 is the unique torsion free connection on (N, g_0) satisfying

$$X(g_0(Y, Z)) = g_0(\nabla^0_X Y, Z) + g_0(Y, \nabla^0_X Z)$$

for all vector fields X, Y, Z on N. The second fundamental form, which measures the difference between ∇ and ∇^0, may be defined just as for Riemannian submanifolds [cf. Hermann (1968, p. 319), Bölts (1977, p. 25, pp. 51–52)].

Definition 3.48. (*Second Fundamental Form*) Let N be a nondegenerate submanifold of (M, g). Given $n \in T_p^{\perp}N$, define the *second fundamental* *form* $S_n : T_pN \times T_pN \to \mathbb{R}$ *in the direction* n as follows. Given $x, y \in T_pN$, extend to local vector fields X, Y tangent to N, and put

$$S_n(x,y) = g\left(\nabla_X Y|_p, n\right) = g\left(\nabla_X Y|_p - \nabla^0_X Y\big|_p, n\right).$$

Define the *second fundamental form* $S : T_p^\perp N \times T_pN \times T_pN \to \mathbb{R}$ by

$$S(n,x,y) = S_n(x,y).$$

Given $n \in T_p^\perp N$, the *second fundamental form operator* $L_n : T_pN \to T_pN$ is defined by $g(L_n(x), y) = S_n(x,y)$ for all $x, y \in T_pN$.

It may be checked that this definition of $S_n(x,y)$ is independent of the choice of extensions X, Y for x, $y \in T_pN$ and also that $S_n : T_pN \times T_pN \to \mathbb{R}$ is a symmetric bilinear map. Furthermore, $S : T_p^\perp N \times T_pN \times T_pN \to \mathbb{R}$ is trilinear for each $p \in N$.

Lemma 3.49. *Let N be a nondegenerate submanifold of (M,g). The second fundamental form $S = 0$ on N iff $\nabla_X Y = \nabla^0_X Y$ for all vector fields X and Y tangent to N.*

Proof. Obviously, Definition 3.48 implies that if $\nabla_X Y = \nabla^0_X Y$ for all vector fields tangent to N, then $S = 0$.

Now suppose $S = 0$. Let $p \in N$ be an arbitrary point. We then have $g(\nabla_X Y|_p - \nabla^0_X Y\big|_p, n) = 0$ for all $n \in T_p^\perp N$ and vector fields X, Y tangent to N. Since $g \mid T_pN \times T_pN$ is nondegenerate, $g \mid T_p^\perp N \times T_p^\perp N$ is also nondegenerate. Thus $\nabla_X Y|_p$ and $\nabla^0_X Y\big|_p$ have the same projection onto $T_p^\perp N$. Since $T_pM = T_pN \oplus T_p^\perp N$, we have $\nabla_X Y|_p = \nabla^0_X Y\big|_p$, as required. □

The second fundamental form may be used to characterize totally geodesic nondegenerate submanifolds of (M,g). A submanifold N of (M,g) is said to be *geodesic* at $p \in N$ if each geodesic γ of (M,g) with $\gamma(0) = p$ and $\gamma'(0) \in T_pN$ is contained in N in some neighborhood of p. The submanifold N is said to be *totally geodesic* if it is geodesic at each of its points. The following proposition is the Lorentzian analogue of a well-known Riemannian result [cf. Hermann (1968, p. 338), Cheeger and Ebin (1975, p. 23)].

Proposition 3.50. *Let N be a nondegenerate submanifold of (M,g). Then N is totally geodesic iff the second fundamental form S satisfies $S=0$ on N.*

Proof. Given that $S=0$ on N, Lemma 3.49 implies that $\nabla_X Y = \nabla^0_X Y$ for all vector fields X, Y tangent to N. Let $c:(-\epsilon,\epsilon)\to M$ be a geodesic in (M,g) with $c'(0)=v\in T_pN$ for some $p\in N$. Also let $\gamma:(-\delta,\delta)\to N$ be the geodesic in (N,g_0) with $\gamma'(0)=v$. Since $\nabla_{\gamma'}\gamma' = \nabla^0_{\gamma'}\gamma' = 0$, the curve γ is also a geodesic in (M,g). Set $\eta=\min\{\epsilon,\delta\}$. From the uniqueness of the geodesic in (M,g) with the given initial direction v, we have $c(t)=\gamma(t)$ for all $t\in(-\eta,\eta)$. Hence $\gamma\,|\,(-\eta,\eta)\subseteq N$ as required.

Conversely, suppose N is totally geodesic in (M,g). Let $p\in N$ be arbitrary. Given $n\in T_p^{\perp}N$ and $x\in T_pN$, let $c:J\to N$ be the geodesic (in both M and N) with $c'(0)=x$. Extend $c'(t)$ to a vector field X tangent to N near p. We then have $S(n,x,x)=g(\nabla_X X|_p, n)=g(\nabla_{c'}c'(0),n)=g(0,n)=0$. By polarization, it follows that $S(n,x,y)=0$ for all $x,y\in T_pN$. Hence $S=0$ on N. $\square$

As will be seen in Chapter 12, the second fundamental form plays an important role in singularity theory in general relativity.

3.6 Warped Products

If (M,g) and (H,h) are two Riemannian manifolds, there is a natural product metric g_0 defined on the product manifold $M\times H$ such that $(M\times H, g_0)$ is again a Riemannian manifold. Bishop and O'Neill (1969) studied a larger class of Riemannian manifolds, including products, which they called *warped products*. If (M,g) and (H,h) are two Riemannian manifolds and $f:M\to(0,\infty)$ is any smooth function, the product manifold $M\times H$ equipped with the metric $g\oplus fh$ is said to be a *warped product* and $f:M\to(0,\infty)$ is called the *warping function*. Following Bishop and O'Neill, we will denote the Riemannian manifold $(M\times H, g\oplus fh)$ by $M\times_f H$. Bishop and O'Neill (1969, p. 23) showed that $M\times_f H$ is a complete Riemannian manifold if and only if both (M,g) and (H,h) are complete Riemannian manifolds. Utilizing this result, they were able to construct a wide variety of complete Riemannian manifolds of everywhere negative sectional curvature using warped products.

In this section, we will use warped product metrics to construct Lorentzian manifolds and will then study the causal structure and completeness properties of this class of Lorentzian manifolds. The theory for Lorentzian manifolds differs from the Riemannian theory somewhat, since the product of two Lorentzian manifolds (M, g) and (H, h) has signature $(-, -, +, \cdots, +)$ and hence is not Lorentzian. Nevertheless, warped product Lorentzian metrics may be constructed from products of Lorentzian and Riemannian manifolds. In particular, this product construction may be used to produce examples of bi-invariant Lorentzian metrics for Lie groups (cf. Section 5.5). A treatment of warped products of semi-Riemannian (not necessarily Lorentzian) manifolds, including a calculation of their Riemannian and Ricci curvature tensors, is given in O'Neill (1983).

Throughout this section, we will let $\pi : M \times H \to M$ and $\eta : M \times H \to H$ denote the projection maps given by $\pi(m, h) = m$ and $\eta(m, h) = h$ for $(m, h) \in M \times H$.

Definition 3.51. (*Lorentzian Warped Product*) Let (M, g) be an n-dimensional manifold ($n \geq 1$) with a signature of $(-, +, \cdots, +)$, and let (H, h) be a Riemannian manifold. Let $f : M \to (0, \infty)$ be a smooth function. The *Lorentzian warped product* $M \times_f H$ is the manifold $\overline{M} = M \times H$ equipped with the Lorentzian metric $\overline{g}$ defined for $v, w \in T_{\overline{p}}\overline{M}$ by

$$\overline{g}(v, w) = g(\pi_* v, \pi_* w) + f(\pi(\overline{p})) \cdot h(\eta_* v, \eta_* w).$$

Definition 3.52. (*Lorentzian Product*) A warped product $M \times_f H$ with $f = 1$ is said to be a *Lorentzian product* and will be denoted by $M \times H$.

Remark 3.53. One may also obtain Lorentzian manifolds by considering warped products $H \times_f M$, where (H, h) is a Riemannian manifold, (M, g) is a Lorentzian manifold, and $f : H \to (0, \infty)$ is a smooth function. The universal covering manifold of anti-de Sitter space (cf. Section 5.3) is an example of a space–time important in general relativity which may be written as a warped product of the form $H \times_f M$ with H Riemannian and M Lorentzian but *not* as a warped product of the form $M \times_f H$ of Definition 3.51. We will only treat warped products of the form $M \times_f H$ in this section.

We begin our study of the causal properties of warped products with the following lemma.

Lemma 3.54. *The warped product $M \times_f H$ of (M, g) and (H, h) may be time oriented iff either (M, g) is time oriented (if $\dim M \geq 2$) or (M, g) is a one-dimensional manifold with a negative definite metric.*

Proof. Suppose that $M \times_f H$ is time orientable. If $\dim M = 1$, then (M, g) has a negative definite metric by Definition 3.51. Now consider the case $\dim M \geq 2$. Since $M \times_f H$ is time orientable, there exists a continuous timelike vector field X for $M \times_f H$. Since $f > 0$ and h is positive definite, we then have $g(\pi_* X, \pi_* X) \leq \overline{g}(X, X) < 0$. Thus the vector field $\pi_* X$ provides a time orientation for (M, g).

Conversely, suppose first that $\dim M \geq 2$ and (M, g) is time oriented by the timelike vector field V. Then V may be lifted to a timelike vector field $\overline{V}$ on $M \times H$ which satisfies $\pi_* \overline{V} = V$ and $\eta_* \overline{V} = 0$. Explicitly, fixing $\overline{p} = (m, b) \in M \times H$, there is a natural isomorphism

$$T_{\overline{p}}(M \times_f H) = T_{\overline{p}}(M \times H) \cong T_m M \times T_b H.$$

Thus we may define $\overline{V}$ at $\overline{p}$ by setting $\overline{V}(\overline{p}) = (V(m), 0_b)$ using this isomorphism to identify $T_{\overline{p}}(M \times H)$ and $T_m M \times T_b H$. It is immediate from Definition 3.51 that $\overline{g}(\overline{V}, \overline{V}) = g(V, V) < 0$. Hence $\overline{V}$ time orients $M \times_f H$ as required.

Now consider the case $\dim M = 1$. It is then known that M is diffeomorphic to S^1 or $\mathbb{R}$. In either case, let T be a smooth vector field on M with $g(T, T) = -1$. Defining $\overline{T}(\overline{p}) = (T(\pi(\overline{p})), 0_{\eta(\overline{p})})$ as above, we have $\eta_* \overline{T} = 0$, so that $\overline{T}$ time orients $\overline{M}$. Note also in the case that $M = S^1$ the integral curves of $\overline{T}$ in $\overline{M}$ are closed timelike curves. Thus $\overline{M}$ is not chronological. □

Lemma 3.55. *Let (H, h) be an arbitrary Riemannian manifold, and let $M = (a, b)$ with $-\infty \leq a < b \leq +\infty$ be given the negative definite metric $-dt^2$. For any smooth function $f : M \to (0, \infty)$, the warped product $(M \times_f H, \overline{g})$ is stably causal.*

Proof. The projection map $\pi : M \times H \to M \subseteq \mathbb{R}$ serves as a time function. □

From the hierarchy of causality conditions given in Figure 3.3, we then obtain

Corollary 3.56. *Let (H, h) be an arbitrary Riemannian manifold, and let $M = (a, b)$ with $-\infty \leq a < b \leq +\infty$ be given the negative definite metric $-dt^2$. For any smooth function $f : M \to (0, \infty)$, the warped product $(M \times_f H, \overline{g})$ is chronological, causal, distinguishing, and strongly causal.*

In the proof of Lemma 3.54 above, we have seen that if $M = S^1$, then the warped product $(S^1 \times_f H, \overline{g})$ fails to be chronological and hence fails to be causal, distinguishing, or strongly causal.

We now list some elementary properties of warped products that follow directly from Definition 3.51. A *homothetic map* $F : (M_1, g_1) \to (M_2, g_2)$ is a diffeomorphism such that $F^*(g_2) = c\, g_1$ for some constant c. We remark that some authors only require homothetic maps to be smooth and not necessarily one-to-one.

Remark 3.57. Let $M \times_f H$ be a Lorentzian warped product.

(1) For each $b \in H$, the restriction $\pi|_{\eta^{-1}(b)} : \eta^{-1}(b) \to M$ is an isometry of $\eta^{-1}(b)$ onto M.

(2) For each $m \in M$, the restriction $\eta|_{\pi^{-1}(m)} : \pi^{-1}(m) \to H$ is a homothetic map of $\pi^{-1}(m)$ with homothetic factor $1/f(m)$.

(3) If $v \in T(M \times H)$, then $g(\pi_* v, \pi_* v) \leq \overline{g}(v, v)$. Thus $\pi_* : T_p(M \times H) \to T_{\pi(p)}M$ maps nonspacelike vectors to nonspacelike vectors, and π maps nonspacelike curves of $M \times_f H$ to nonspacelike curves of M.

(4) Since $|g(\pi_* v, \pi_* v)| \geq |\overline{g}(v, v)|$ if $v \in T(M \times H)$ is nonspacelike, the map π is length nondecreasing on nonspacelike curves (cf. Section 4.1, formula (4.1) for the definition of Lorentzian arc length).

(5) For each $(m, b) \in M \times H$, the submanifolds $\pi^{-1}(m)$ and $\eta^{-1}(b)$ of $M \times_f H$ are nondegenerate in the sense of Definition 3.47.

(6) If $\phi : H \to H$ is an isometry, then the map $\Phi = 1 \times \phi : M \times_f H \to M \times_f H$ given by $\Phi(m, b) = (m, \phi(b))$ is an isometry of $M \times_f H$.

(7) If $\psi : M \to M$ is an isometry of M such that $f \circ \psi = f$, then the map $\Psi = \psi \times 1 : M \times_f H \to M \times_f H$ given by $\Psi(m, b) = (\psi(m), b)$ is an

isometry of $M \times_f H$. Thus if X is a Killing vector field on M (i.e., $L_X g = 0$) with $X(f) = 0$, then the natural lift $\overline{X}$ of X to $M \times_f H$ given by $\overline{X}(p) = (X(\pi(p)), 0_{\eta(p)})$ is a Killing vector field on $M \times_f H$.

Lemma 3.58. *Let $M \times_f H$ be a Lorentzian warped product. Then for each $b \in H$, the leaf $\eta^{-1}(b)$ is totally geodesic.*

Proof. Since the map $\pi : M \times_f H \to M$ is length nondecreasing on nonspacelike curves and since nonspacelike geodesics are locally length maximizing, it follows that any nonspacelike geodesic of $\eta^{-1}(b)$ (in the metric induced by the inclusion $\eta^{-1}(b) \subseteq M \times_f H$) is a geodesic in the ambient manifold $M \times_f H$. Thus the second fundamental form vanishes on all nonspacelike vectors in $T(\eta^{-1}(b))$. Since any tangent vector in $T(\eta^{-1}(b))$ may be written as a linear combination of nonspacelike vectors in $T(\eta^{-1}(b))$, it follows that the second fundamental form vanishes identically. Hence, $\eta^{-1}(b)$ is totally geodesic by Proposition 3.50. □

In view of Corollary 3.56, we may now restrict our attention to studying the fundamental causal properties of time oriented Lorentzian warped products $(M \times_f H, \overline{g})$ with $\dim M \geq 2$.

Lemma 3.59. *Let $p = (p_1, p_2)$ and $q = (q_1, q_2)$ be two points in $M \times_f H$ with $p \ll q$ (respectively, $p \leqslant q$) in $(M \times_f H, \overline{g})$. Then $p_1 \ll q_1$ (respectively, $p_1 \leqslant q_1$) in (M, g).*

Proof. If γ is a future directed timelike (respectively, nonspacelike) curve in $M \times_f H$ from p to q, then $\pi \circ \gamma$ is a future directed timelike (respectively, nonspacelike) curve in M from p_1 to q_1. □

While $\pi : M \times_f H \to M$ takes nonspacelike curves to nonspacelike curves, π does *not* preserve null curves. Indeed, it follows from Definition 3.51 that if γ is any smooth null curve with $\eta_* \gamma(t) \neq 0$ for all t, then $g(\pi_* \gamma(t), \pi_* \gamma(t)) < 0$ for all t.

For points p and q in the same leaf $\eta^{-1}(b)$ of $M \times_f H$, Lemma 3.59 may be strengthened as follows.

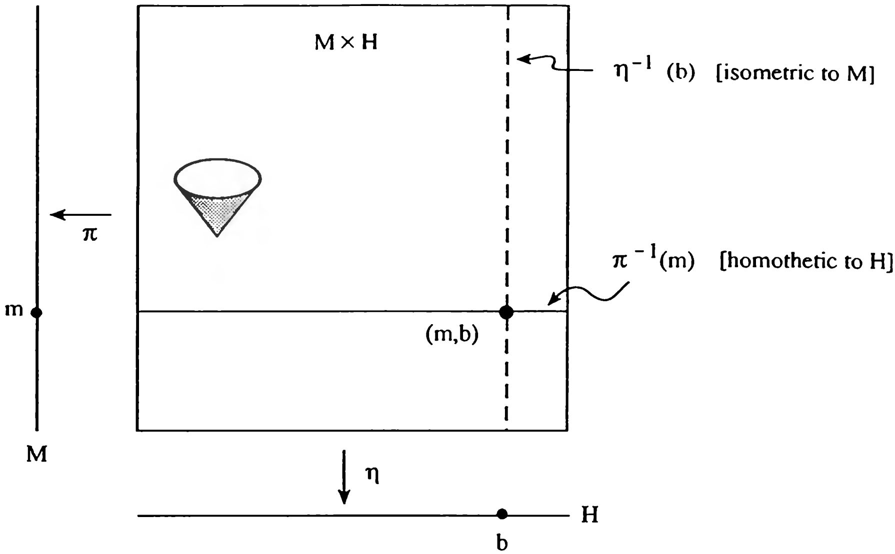


FIGURE 3.9. Let (m, b) be a point of the warped product $M \times_f H$. Then the projection map π restricted to $\eta^{-1}(b)$ is an isometry onto M, and the projection map η restricted to $\pi^{-1}(m)$ is a homothetic map onto H.

Lemma 3.60. *If $p = (p_1, b)$ and $q = (q_1, b)$ are points in the same leaf $\eta^{-1}(b)$ of $M \times_f H$, then $p \ll q$ (respectively, $p \leqslant q$) in $(M \times_f H, \overline{g})$ iff $p_1 \ll q_1$ (respectively, $p_1 \leqslant q_1$) in (M, g).*

Proof. By Lemma 3.59, it only remains to show that if $p_1 \ll q_1$ (respectively, $p_1 \leqslant q_1$) in (M, g), then $p \ll q$ (respectively, $p \leqslant q$) in $(M \times_f H, \overline{g})$. But if $\gamma_1 : [0, 1] \to M$ is a future directed timelike (respectively, nonspacelike) curve in M from p_1 to q_1, then $\gamma(t) = (\gamma_1(t), b)$, $0 \leq t \leq 1$, is a future directed timelike (respectively, nonspacelike) curve in $M \times_f H$ from p to q. $\square$

Lemma 3.60 implies that each leaf $\eta^{-1}(b)$, $b \in H$, has the same chronology and causality as (M, g). In particular, Lemmas 3.59 and 3.60 imply that $(M \times_f H, \overline{g})$ has a closed timelike (respectively, nonspacelike) curve iff (M, g) has a closed timelike (respectively, nonspacelike) curve. Hence

Proposition 3.61. *Let (M, g) be a space–time, and let (H, h) be a Riemannian manifold. Then the Lorentzian warped product $(M \times_f H, \overline{g})$ is chronological (respectively, causal) iff (M, g) is chronological (respectively, causal).*

A similar result holds for strong causality.

Proposition 3.62. *Let (M, g) be a space–time, and let (H, h) be a Riemannian manifold. Then the Lorentzian warped product $(M \times_f H, \overline{g})$ is strongly causal iff (M, g) is strongly causal.*

Proof. We first show that if the space–time (M, g) is not strongly causal at p_1, then $(M \times_f H, \overline{g})$ is not strongly causal at $p = (p_1, b)$ for any $b \in H$. Since (M, g) is not strongly causal at p_1, there is an open neighborhood U_1 of p_1 in M and a sequence $\{\gamma_k : [0, 1] \to M\}$ of future directed nonspacelike curves with $\gamma_k(0) \to p_1$ and $\gamma_k(1) \to p_1$ as $k \to \infty$, but $\gamma_k(1/2) \notin U_1$ for all k. Define $\sigma_k : [0, 1] \to M \times H$ by $\sigma_k(t) = (\gamma_k(t), b)$. Let V_1 be any open neighborhood of b in H, and set $U = U_1 \times V_1$ in $M \times H$. Then U is an open neighborhood of $p = (p_1, b)$ in $M \times_f H$, and $\{\sigma_k\}$ is a sequence of nonspacelike future directed curves in $M \times_f H$ with $\sigma_k(0) \to p$ and $\sigma_k(1) \to p$ as $k \to \infty$, but $\sigma_k(1/2) \notin U$ for all k. Thus $(M \times_f H, \overline{g})$ is not strongly causal at p.

Conversely, suppose that strong causality fails at the point $p = (p_1, q_1)$ of $(M \times_f H, \overline{g})$. Let $(x_1, \ldots, x_i)$ be local coordinates on M near p_1 such that g has the form $\operatorname{diag}(-1, +1, \ldots, +1)$ at p_1, and let $(x_{i+1}, \ldots, x_n)$ be local coordinates on H near q_1 such that fh has the form $\operatorname{diag}(+1, \ldots, +1)$ at q_1. Then $(x_1, \ldots, x_i, x_{i+1}, \ldots, x_n)$ are local coordinates for $M \times_f H$ near p. Furthermore, $F_1 = x_1$ and $F_2 = x_1 \circ \pi$ are (locally defined) time functions for M near p_1 and for $M \times_f H$ near p, respectively. The failure of strong causality at p implies the existence of a sequence $\gamma_k : [0, 1] \to M \times_f H$ of future directed nonspacelike curves with $\gamma_k(0) \to p$ and $\gamma_k(1) \to p$ as $k \to \infty$, but

$F_2(\gamma_k(1/2)) \geq \epsilon > 0$ for all k and some parametrization of γ_k. Choose a neighborhood W of p_1 in M such that W is covered by the local coordinates $(x_1, \ldots, x_i)$ above and such that $\sup\{F_1(r) : r \in W\} \leq \epsilon/2$. The curves $\pi \circ \gamma_k$ are then future directed nonspacelike curves in M with $\pi \circ \gamma_k(0) \to p_1$, $\pi \circ \gamma_k(1) \to p_1$, and $\pi \circ \gamma_k(1/2) \notin W$. Hence W and $\{\pi \circ \gamma_k\}$ show that strong causality fails at p_1 in (M, g) as required. □

In Proposition 3.64 we prove the equivalence of stable causality for (M, g) and $(M \times_f H, \overline{g})$ for $\dim M \geq 2$. From this proposition and the last two propositions, it follows that the basic causal properties of $(M \times_f H, \overline{g})$ are determined by those of (M, g).

Remark 3.63. If $g < g_1$ on M, then there is a smooth conformal factor $\Omega : M \to (0, \infty)$ such that $\Omega g_1(v, v) < g(v, v)$ for all nontrivial vectors which are nonspacelike with respect to g.

Proposition 3.64. *Let (M, g) be a space–time and (H, h) a Riemannian manifold. Then the Lorentzian warped product $(M \times_f H, \overline{g})$ is stably causal iff (M, g) is stably causal.*

Proof. In this proof we will use the identification $T_p(M \times H) \cong T_{p_1}M \times T_bH$ for all $p = (p_1, b) \in M \times H$.

Assuming that $(M \times_f H, \overline{g})$ is stably causal, there exists $\overline{g}_1 \in \mathrm{Lor}(M \times H)$ such that $\overline{g} < \overline{g}_1$ and $\overline{g}_1$ is causal. If b is a fixed point of H, then we may assume without loss of generality that $\overline{g}_1|_{\eta^{-1}(b)}$ is nondegenerate since $\overline{g}|_{\eta^{-1}(b)}$ is nondegenerate. Setting $\widetilde{g}_1 = \overline{g}_1|_{\eta^{-1}(b)}$ and using $\pi|_{\eta^{-1}(b)}$ to identify $\eta^{-1}(b)$ with M, we obtain a metric $g_1 \in \mathrm{Lor}(M)$ such that $\pi|_{\eta^{-1}(b)}$ is an isometry of $(\eta^{-1}(b), \widetilde{g}_1)$ onto (M, g_1). Notice that since $(M \times H, \overline{g}_1)$ is causal, the space–time $(\eta^{-1}(b), \widetilde{g}_1)$ is causal, and hence (M, g_1) is also causal. To show $g < g_1$ on M, we choose a nonzero vector $v_1 \in T_{p_1}M$ such that $g(v_1, v_1) \leq 0$. If 0_b denotes the zero vector in T_bH, then $\overline{g}(v, v) = g(v_1, v_1) \leq 0$, where $v = (v_1, 0_b) \in T_{p_1}M \times T_bH$. Since $\overline{g} < \overline{g}_1$, we obtain $\overline{g}_1(v, v) = g_1(v_1, v_1) < 0$. Hence $g < g_1$, and (M, g) is stably causal.

Conversely, we now assume that (M, g) is stably causal. Let $g_1 \in \mathrm{Lor}(M)$ be a causal metric with $g < g_1$. By Remark 3.63 we may also assume that

$g_1(v_1, v_1) < g(v_1, v_1)$ for all vectors $v_1 \neq 0$ which are nonspacelike with respect to g. Since Proposition 3.61 implies that $\overline{g}_1 = g_1 \oplus fh$ is a causal metric on $M \times H$, it suffices to show that $\overline{g} < \overline{g}_1$. To this end, let $v = (v_1, v_2)$ be a nontrivial vector of $T_p(M \times H)$ which is nonspacelike with respect to $\overline{g}$. Then since $\overline{g}(v,v) = g(v_1, v_1) + f(\pi(v)) \cdot h(v_2, v_2) \leq 0$ and $f(\pi(v)) \cdot h(v_2, v_2) > 0$ with $v_2 \neq 0$, the nontriviality of v implies that $v_1 \neq 0$ and $g(v_1, v_1) \leq 0$. Thus $\overline{g}_1(v,v) = g_1(v_1, v_1) + f(\pi(v)) \cdot h(v_2, v_2) < g(v_1, v_1) + f(\pi(v)) \cdot h(v_2, v_2) \leq 0$ which shows that $\overline{g} < \overline{g}_1$ and establishes the proposition. □

Geroch's Splitting Theorem (cf. Theorem 3.17) guarantees that any globally hyperbolic space–time may be written as a topological product $\mathbb{R} \times S$ where S is a Cauchy hypersurface. Geroch's result suggests investigating conditions on (M, g) and (H, h) which imply that the warped product $(M \times_f H, \overline{g})$ is globally hyperbolic. These conditions are given for $\dim M = 1$ and $\dim M \geq 2$ in Theorems 3.66 and 3.68, respectively. In order to prove these results, it is first necessary to show that a curve in a complete Riemannian manifold which is inextendible in one direction must have infinite length.

Lemma 3.65. *Let (H, h) be a complete Riemannian manifold, and let $\gamma : [0, 1) \to H$ be a curve of finite length in (H, h). Then there exists a point $p \in H$ such that $\gamma(t) \to p$ as $t \to 1^-$.*

Proof. Let d_0 denote the Riemannian distance function induced on H by the Riemannian metric h. Let $L = L_0(\gamma)$ be the Riemannian arc length of γ, and set $K = \{q \in H : d_0(\gamma(0), q) \leq L\}$. The Hopf–Rinow Theorem [cf. Hicks (1965, pp. 163–164)] implies that K is compact. Fix a sequence $\{t_n\}$ in $[0, 1)$ with $t_n \to 1$. Since $d(\gamma(0), \gamma(t)) \leq L(\gamma| \, [0, t]) \leq L$ for $t \in [0, 1)$, we have $\gamma[0, 1) \subseteq K$. Thus by the compactness of K, the sequence $\{\gamma(t_n)\}$ has a limit point $p \in K$. If $\lim_{t \to 1^-} \gamma(t) \neq p$, there would then exist an $\epsilon > 0$ such that γ leaves the ball $\{m \in M : d(p, m) \leq \epsilon\}$ infinitely often. But this would imply that γ has infinite length, in contradiction. □

The following theorem may be obtained from the combination of Corollary 3.56 and Lemma 3.65. The proof, which is similar to that of Theorem 3.68, will be omitted.

Theorem 3.66. *Let (H,h) be a Riemannian manifold, and let $M = (a,b)$ with $-\infty \le a < b \le +\infty$ be given the negative definite metric $-dt^2$. Then the Lorentzian warped product $(M \times_f H, \overline{g})$ is globally hyperbolic iff (H,h) is complete.*

Theorem 3.66 may be regarded as a "metric converse" to Geroch's splitting theorem. If $f = 1$ is assumed, so that the warped product $(M \times_f H, \overline{g})$ is simply a metric product $(M \times H, g \oplus h)$, Theorem 3.66 may be strengthened to include geodesic completeness (cf. Definition 6.2 for the definition of geodesic completeness).

Theorem 3.67. *Suppose that (H,h) is a Riemannian manifold and that $\mathbb{R} \times H$ is given the product Lorentzian metric $-dt^2 \oplus h$. Then the following are equivalent:*

(1) *(H,h) is geodesically complete.*
(2) *$(\mathbb{R} \times H, -dt^2 \oplus h)$ is geodesically complete.*
(3) *$(\mathbb{R} \times H, -dt^2 \oplus h)$ is globally hyperbolic.*

Proof. We know that (1) iff (3) from Theorem 3.66. Thus it remains to show (1) iff (2). But this is a consequence of the fact that all geodesics of $\mathbb{R} \times H$ are either (up to parametrization) of the form $(\lambda t, c(t))$, $(\lambda_0, c(t))$, or $(\lambda t, h_0)$, where λ, $\lambda_0 \in \mathbb{R}$ are constants, $h_0 \in H$, and $c : J \to H$ is a unit speed geodesic in H. $\square$

Suppose that a space–time (M,g) of dimension $n \ge 3$ satisfies the timelike convergence condition (i.e., has everywhere nonnegative nonspacelike Ricci curvatures) and satisfies the generic condition (i.e., each inextendible nonspacelike geodesic contains a point where the tangent vector W satisfies the equation $\sum_{c,d=1}^{n} W^c W^d W_{[a} R_{b]cd[e} W_{f]} \ne 0$ [cf. Chapter 2]). Then if (M,g) has a compact Cauchy surface, the space–time (M,g) is geodesically incomplete. Thus Theorem 3.66 may *not* be strengthened for arbitrary warped products to include geodesic completeness. The "big bang" Robertson–Walker cosmological models (cf. Section 5.4) are examples of globally hyperbolic warped products which are not geodesically complete.

In contrast, let $(\mathbb{R} \times H, -dt^2 \oplus h)$ be a product space–time of the form considered in Theorem 3.67. Fix any $b_0 \in H$. Then $\gamma(t) = (t, b_0)$ is a timelike geodesic with $R(\gamma'(t), v) = 0$ for all $v \in T_{\gamma(t)}(\mathbb{R} \times H)$ for each $t \in \mathbb{R}$. Thus $(\mathbb{R} \times H, -dt^2 \oplus h)$ *fails* to satisfy the generic condition.

If $\dim M = 1$ and M is homeomorphic to $\mathbb{R}$, we have just given necessary and sufficient conditions for the warped product $M \times_f H$ to be globally hyperbolic. If $M = S^1$, we remarked above that $(M \times_f H, \overline{g})$ is nonchronological no matter which Riemannian metric h is chosen for H. Thus no warped product space–time $(S^1 \times_f H, \overline{g})$ is globally hyperbolic.

We now consider the case $\dim M \geq 2$.

Theorem 3.68. *Let (M, g) be a space–time, and let (H, h) be a Riemannian manifold. Then the Lorentzian warped product $(M \times_f H, \overline{g})$ is globally hyperbolic iff both of the following conditions are satisfied:*

(1) *(M, g) is globally hyperbolic.*

(2) *(H, h) is a complete Riemannian manifold.*

Proof. ($\Rightarrow$) Suppose first that $(M \times_f H, \overline{g})$ is globally hyperbolic. Fixing $b \in H$, we may identify (M, g) with the closed submanifold $\eta^{-1}(b) = M \times \{b\}$ since the projection map $\pi : \eta^{-1}(b) \to M$ is an isometry. Lemma 3.60 implies that under this identification, the set $J^+(p_1) \cap J^-(q_1)$ in M corresponds to $\eta^{-1}(b) \cap J^+((p_1, b)) \cap J^-((q_1, b))$ in $(M \times_f H)$ for any p_1 and q_1 in M. Since $\eta^{-1}(b)$ is closed and $(M \times_f H, \overline{g})$ is globally hyperbolic, $\eta^{-1}(b) \cap J^+((p_1, b)) \cap J^-((q_1, b))$ is compact in $(M \times_f H)$. Hence $J^+(p_1) \cap J^-(q_1)$ is compact in M. Because $(M \times_f H, \overline{g})$ is globally hyperbolic, it is also strongly causal. Thus (M, g) is strongly causal by Proposition 3.62. Hence (M, g) is globally hyperbolic as required.

Now we show that $(M \times_f H, \overline{g})$ globally hyperbolic implies that (H, h) is a complete Riemannian manifold. We will suppose that (H, h) is incomplete and derive a contradiction to the global hyperbolicity of $(M \times_f H, \overline{g})$. For this purpose, fix any pair of points p_1 and q_1 in M with $p_1 \ll q_1$, and let $\gamma_1 : [0, L] \to M$ be a unit speed future directed timelike curve in M from p_1 to q_1. Set $\alpha = \sup\{f(\gamma_1(t)) : t \in [0, L]\}$ where $f : M \to (0, \infty)$ is the

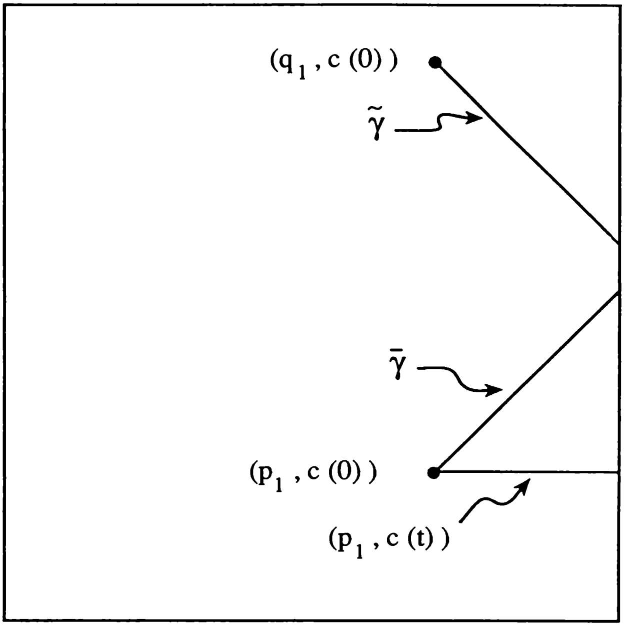


FIGURE 3.10. In the proof of Theorem 3.68, the curve $c : [0, \beta) \to H$ is a geodesic which is not extendible to $t = \beta < \infty$. The curves $\overline{\gamma}(t) = (\gamma_1(t), c(t))$ and $\widetilde{\gamma}(t) = (\gamma_1(L - t), c(t))$ are inextendible nonspacelike curves in $(M \times_f H, \overline{g}\,)$ and hence do not have compact closure.

given warping function. Since $\gamma_1([0, L])$ is a compact subset of M, we have $0 < \alpha < \infty$.

Assuming that (H, h) is not complete, the Hopf–Rinow Theorem ensures the existence of a geodesic $c : [0, \beta) \to H$ with $h(c'(t), c'(t)) = 1/\alpha$ which is not extendible to $t = \beta < \infty$. By changing $c(0)$ and reparametrizing c if necessary, we may suppose that $0 < \beta < L/2$. Define a future directed nonspacelike curve $\overline{\gamma} : [0, \beta) \to M \times H$ and a past directed nonspacelike curve $\widetilde{\gamma} : [0, \beta) \to M \times H$ by $\overline{\gamma}(t) = (\gamma_1(t), c(t))$ and $\widetilde{\gamma}(t) = (\gamma_1(L - t), c(t))$, respectively. For each t with $0 \leq t < \beta$, we have $\gamma_1(t) \ll \gamma_1(L - t)$ in (M, g) since $t < L - t$. Hence by Lemma 3.60, we have $(\gamma_1(t), c(t)) \ll (\gamma_1(L-t), c(t))$ in $M \times_f H$. Thus $(p_1, c(1)) \leqslant \overline{\gamma}(t) \ll \widetilde{\gamma}(t) \leqslant (q_1, c(0))$ for all $0 \leq t < \beta$ (cf. Figure 3.10). It follows that $\overline{\gamma}([0,\beta))$ is contained in $J^+\big((p_1, c(0))\big) \cap J^-\big((q_1, c(0))\big)$. Since $c = \eta \circ \overline{\gamma}$ does not have compact closure in H, the curve $\overline{\gamma} : [0, \beta) \to M \times_f H$ does not have compact closure in $J^+\big((p_1, c(0))\big) \cap J^-\big((q_1, c(0))\big)$. But since $(M \times_f H, \overline{g})$ is globally hyperbolic, the set $J^+\big((p_1, c(0))\big) \cap J^-\big((q_1, c(0))\big)$ is compact, in contradiction.

($\Leftarrow$) Suppose now that (M, g) is globally hyperbolic. Assuming that the warped product $(M \times_f H, \overline{g})$ is not globally hyperbolic, we must show that (H, h) is not complete. Since (M, g) is strongly causal, $(M \times_f H, \overline{g})$ is also strongly causal by Proposition 3.62. Hence since $(M \times_f H, \overline{g})$ is not globally hyperbolic, there exist distinct points (p_1, b_1) and (p_2, b_2) in $M \times_f H$ such that $J^+((p_1, b_1)) \cap J^-((p_2, b_2))$ is noncompact. There is then a future directed nonspacelike curve $\gamma : [0, 1) \to J^+((p_1, b_1)) \cap J^-((p_2, b_2))$ which is future inextendible in $(M \times_f H, \overline{g})$. Let $\gamma(t) = (u_1(t), u_2(t))$, where $u_1 : [0, 1) \to M$ and $u_2 : [0, 1) \to H$. Then $u_1 : [0, 1) \to M$ is a future directed nonspacelike curve contained in $J^+(p_1) \cap J^-(p_2)$. Since (M, g) is globally hyperbolic, $J^+(p_1) \cap J^-(p_2)$ is compact. Hence if we set $\alpha_0 = \inf\{f(m) : m \in J^+(p_1) \cap J^-(p_2)\}$, then $\alpha_0 > 0$. Also since (M, g) is strongly causal, no future directed, future inextendible, nonspacelike curve may be future imprisoned in the compact set $J^+(p_1) \cap J^-(p_2)$ (cf. Proposition 3.13). Hence there exists a point $r \in J^+(p_1) \cap J^-(p_2)$ with $\lim_{t \to 1^-} u_1(t) = r$. We may then extend u_1 to a continuous curve $u_1 : [0, 1] \to M$ by setting $u_1(1) = r$. Since the curve $\gamma = (u_1, u_2)$ was inextendible to $t = 1$, it follows that $u_2(t)$ cannot converge to any point of H as $t \to 1^-$. By Lemma 3.65, either (H, h) is incomplete or u_2 has infinite length. As $u_1 : [0, 1] \to M$ is a nonspacelike curve defined on a compact interval, u_1 has finite length in (M, g). Since $f(u_1(t)) \geq \alpha_0 > 0$ for all $t \in [0, 1]$ and

$$\overline{g}(\gamma'(t), \gamma'(t)) = g(u_1{}'(t), u_1{}'(t)) + f(u_1(t)) \cdot h(u_2{}'(t), u_2{}'(t)) \leq 0,$$

it follows that u_2 has finite length in (H, h). Thus (H, h) is incomplete as required. $\square$

Cauchy surfaces may be constructed for globally hyperbolic Lorentzian warped products as follows.

Theorem 3.69. *Let (H, h) be a complete Riemannian manifold. Let $(M \times_f H, \overline{g})$ be the Lorentzian warped product of (M, g) and (H, h).*

(1) *If $M = (a, b)$ with $-\infty \leq a < b \leq +\infty$ is given the metric $-dt^2$, then $\{p_1\} \times H$ is a Cauchy surface of $(M \times_f H, \overline{g})$ for each $p_1 \in M$.*
(2) *If (M, g) is globally hyperbolic with Cauchy surface S_1, then $S_1 \times H$ is a Cauchy surface of $(M \times_f H, \overline{g})$.*

Proof. Since the proofs of (1) and (2) are similar, we shall only give the proof of (2). In this case, $S_1 \times H$ is an achronal subset of $(M \times_f H, \overline{g})$. To show $S_1 \times H$ is a Cauchy surface, we must show that every inextendible nonspacelike curve in $M \times_f H$ meets $S_1 \times H$. Now given $(p_1, p_2) \in (M \times H) - (S_1 \times H)$, either every future directed, future inextendible, nonspacelike curve in (M, g) beginning at p_1 meets S_1 or every past directed, past inextendible, nonspacelike curve starting at p_1 meets S_1. Since the two cases are similar, we will suppose the former holds and then show that every future directed, future inextendible, nonspacelike curve $\gamma : [0, 1) \to M \times_f H$ with $\gamma(0) = (p_1, p_2)$ meets $S_1 \times H$.

Thus suppose that $\gamma : [0, 1) \to M \times_f H$ is a future directed, future inextendible, nonspacelike curve with $\gamma(0) = (p_1, p_2)$ which does not meet $S_1 \times H$. Decompose $\gamma(t) = (u_1(t), u_2(t))$ with $u_1 : [0, 1) \to M$ and $u_2 : [0, 1) \to H$. Since S_1 is a Cauchy surface for (M, g) and (M, g) is globally hyperbolic, the set $J^+(p_1) \cap J^-(S_1)$ is compact [cf. Beem and Ehrlich (1979a, p. 163)]. As in the proof of Theorem 3.68, the strong causality of (M, g) implies that there exists a point $r \in J^+(p_1) \cap J^-(S_1)$ with $\lim_{t \to 1^-} u_1(t) = r$. Since $J^+(p_1) \cap J^-(S_1)$ is compact, the warping function $f : M \to (0, \infty)$ achieves a minimum $\alpha_0 > 0$ on $J^+(p_1) \cap J^-(S_1)$. As in the proof of Theorem 3.68, this then implies that $u_2 : [0, 1) \to H$ has finite length. Since (H, h) is complete, by Lemma 3.65 there exists a point $b \in H$ with $\lim_{t \to 1^-} u_2(t) = b$. Setting $\gamma(1) = (r, b)$, we have then extended γ to a nonspacelike future directed curve $\gamma : [0, 1] \to M \times_f H$, contradicting the inextendibility of γ. Hence γ must meet $S_1 \times H$ as required. $\square$

We now consider the nonspacelike geodesic completeness of the class of Lorentzian warped products of the form $\overline{M} = (a,b) \times_f H$ with $\overline{g} = -dt^2 \oplus fh$. Here a space–time is said to be *null* (respectively, *timelike*) *geodesically incomplete* if some future directed null (respectively, timelike) geodesic cannot be extended to be defined for arbitrary negative and positive values of an affine parameter (cf. Definitions 6.2 and 6.3). Since we are using the metric $-dt^2$ on (a,b), the curve $c(t) = (t, y_0)$ with $y_0 \in H$ fixed is a unit speed timelike geodesic in $(\overline{M}, \overline{g})$ no matter which warping function is chosen. Consequently, if $a > -\infty$ or $b < +\infty$, then $(\overline{M}, \overline{g})$ is timelike geodesically incomplete for all possible warping functions f. Moreover, if a and b are both finite and if γ is any timelike geodesic in $\overline{M} = (a,b) \times_f H$, then $L(\gamma) \leq b - a < \infty$. Thus if a and b are finite, all timelike geodesics are past and future incomplete. Nonetheless, if the warping function f is chosen suitably, $(\overline{M}, \overline{g})$ may be null geodesically complete even if a and b are both finite. This will be clear from the proof of Theorem 3.70 below.

If $\overline{M} = \mathbb{R} \times_f H$ with $\overline{g} = -dt^2 \oplus fh$, then any timelike geodesic of the form $c(t) = (t, y_0)$ is past and future timelike complete. However, warped product space–times $\overline{M} = \mathbb{R} \times_f H$ may be constructed for which all nonspacelike geodesics except for those of the form $t \to (t, y_0)$ are future incomplete. One such example may be given as follows. Busemann and Beem (1966) studied the space–time $\overline{M} = \{(x,y) \in \mathbb{R}^2 : y > 0\}$ with the Lorentzian metric $ds^2 = y^{-2}(dx^2 - dy^2)$. Busemann and Beem (1966, p. 245) noted that all timelike geodesics except for those of the form $t \to (t, y_0)$ are future incomplete. Setting $t = \ln y$, this space–time is transformed into the Lorentzian warped product $\mathbb{R} \times_f \mathbb{R}$ with $\overline{g} = -dt^2 \oplus f dt^2$, where $f(t) = e^{-2t}$. Since the map $F : (\overline{M}, ds^2) \to (\mathbb{R} \times_f \mathbb{R}, \overline{g})$ given by $F(x,y) = (x, \ln y)$ is a global isometry, all timelike geodesics of $(\mathbb{R} \times_f \mathbb{R}, \overline{g})$ except for those of the form $t \to (t, y_0)$ are future incomplete. It will also follow from Theorem 3.70 below that all null geodesics are future incomplete. Similarly, if $(\mathbb{R}^n, h)$ denotes $\mathbb{R}^n$ with the usual Euclidean metric $h = {dx_1}^2 + {dx_2}^2 + \cdots + {dx_n}^2$, then $(\mathbb{R} \times_f \mathbb{R}^n, \overline{g})$ with $\overline{g} = -dt^2 \oplus fh$ and $f(t) = e^{-2t}$ is a space–time with all nonspacelike geodesics future incomplete except for those of the form $t \to (t, y_0)$.

In order to study geodesic completeness, it is necessary to determine the Levi–Civita connection for a Lorentzian warped product metric. For this purpose, we will consider the general warped product $(M \times_f H, g \oplus fh)$ where $f : M \to (0, \infty)$, (H, h) is Riemannian, and (M, g) is equipped with a metric of signature $(-, +, \cdots, +)$. Let ∇^1 denote the Levi–Civita connection for (M, g) and ∇^2 denote the Levi–Civita connection for (H, h). Given vector fields X_1, Y_1 on M and X_2, Y_2 on H, we may lift them to $M \times H$ and obtain the vector fields $X = (X_1, 0) + (0, X_2) = (X_1, X_2)$ and $Y = (Y_1, 0) + (0, Y_2) = (Y_1, Y_2)$ on $M \times H$. Recall that the connection $\overline{\nabla}$ for $(M \times_f H, g \oplus fh)$ is related to the metric $\overline{g} = g \oplus fh$ by the Koszul formula

$$\begin{aligned} 2\overline{g}(\overline{\nabla}_X Y, Z) &= X\,\overline{g}(Y, Z) + Y\,\overline{g}(X, Z) - Z\,\overline{g}(X, Y) \\ &\quad + \overline{g}([X, Y], Z) - \overline{g}([X, Z], Y) - \overline{g}([Y, Z], X) \end{aligned}$$

[cf. Cheeger and Ebin (1975, p. 2)]. Using this formula and setting $\phi = \ln f$, we obtain the following formula for $\overline{\nabla}$ for X and Y as above:

$$\text{(3.17)} \quad \overline{\nabla}_X Y = \nabla^1_{X_1} Y_1 + \nabla^2_{X_2} Y_2 + \frac{1}{2}\left[X_1(\phi) Y_2 + Y_1(\phi) X_2 - \overline{g}(X_2, Y_2)\operatorname{grad}\phi\right].$$

Here $\operatorname{grad}\phi$ denotes the gradient of the function ϕ on (M, g), and we are identifying the vector $\nabla^1_{X_1} Y_1\big|_p \in T_pM$ with the vector $\left(\nabla^1_{X_1} Y_1, 0_q\right) \in T_{(p,q)}(M \times H)$, etc.

We are now ready to obtain the following criterion for null geodesic incompleteness of Lorentzian warped products $\overline{M} = (a, b) \times_f H$ [cf. Beem, Ehrlich, and Powell (1982)]. Throughout the rest of this section, let ω_0 denote an interior point of (a, b).

Theorem 3.70. *Let $\overline{M} = (a, b) \times_f H$ be a Lorentzian warped product with Lorentzian metric $\overline{g} = -dt^2 \oplus fh$ where $-\infty \leq a < b \leq +\infty$, (H, h) is an arbitrary Riemannian manifold, and $f : (a, b) \to (0, \infty)$. Set $S(t) = \sqrt{f(t)}$. Then if $\lim_{t \to a^+} \int_t^{\omega_0} S(s)ds$ $\left[\text{respectively, } \lim_{t \to b^-} \int_{\omega_0}^t S(s)ds\right]$ is finite, every future directed null geodesic in $(\overline{M}, \overline{g})$ is past [respectively, future] incomplete.*

Proof. Let γ_0 be an arbitrary future directed null geodesic in $(\overline{M}, \overline{g})$. We may reparametrize γ_0 to be of the form $\gamma(t) = (t, c(t))$, where γ is a smooth null pregeodesic. Accordingly, there exists a smooth function $g(t)$ such that

$$\overline{\nabla}_{\gamma'}\gamma'\big|_t = g(t)\gamma'(t) = g(t)\left.\frac{\partial}{\partial t}\right|_t + g(t)c'(t)$$

[cf. Hawking and Ellis (1973, p. 33)]. However, since $\gamma'(t) = \partial/\partial t|_t + c'(t)$ and $\overline{g}(\gamma',\gamma') = -1 + \overline{g}(c',c') = 0$, we obtain using formula (3.17) that

$$\begin{aligned}\overline{\nabla}_{\gamma'}\gamma'\big|_t &= \nabla^1_{\frac{\partial}{\partial t}}\left.\frac{\partial}{\partial t}\right|_t + \nabla^2_{c'}c'\big|_t + \tfrac{\partial}{\partial t}(\phi)c'(t) - \tfrac{1}{2}\,\overline{g}(c'(t),c'(t))\operatorname{grad}\phi \\ &= \nabla^2_{c'}c'\big|_t + \frac{f'(t)}{f(t)}c'(t) + \frac{f'(t)}{2f(t)}\left.\frac{\partial}{\partial t}\right|_t.\end{aligned}$$

Equating terms with a $\partial/\partial t$ component, we obtain the formula

$$g(t) = \frac{f'(t)}{2f(t)}. \tag{3.18}$$

Thus $\overline{\nabla}_{\gamma'}\gamma'\big|_t = (1/2)[\ln f(t)]'\,\gamma'(t) = [\ln S(t)]'\,\gamma'(t)$. If we define $p : (a,b) \to \mathbb{R}$ by

$$p(t) = \int_{\omega_0}^{t} S(s)\,ds,$$

then $p'(t) = S(t) > 0$ so that p^{-1} exists. Moreover, from the classical theory of projective transformations we know that the curve $\gamma_1(t) = \gamma \circ p^{-1}(t) = (p^{-1}(t), c \circ p^{-1}(t))$ is a null geodesic [cf. Spivak (1970, pp. 6–35 ff.)]. Let

$$A = \lim_{t\to a^+} p(t) \qquad \text{and} \qquad B = \lim_{t\to b^-} p(t).$$

Since p is monotone increasing, we have that $p : (a,b) \to (A,B)$ is a bijection. Hence $p^{-1} : (A,B) \to (a,b)$ and thus $\gamma_1 = \gamma \circ p^{-1} : (A,B) \to \overline{M}$. Therefore if A is finite, γ_1 is past incomplete, and if B is finite, γ_1 is future incomplete as required. □

It is immediate from Theorem 3.70 that if a and b are finite and the warping function $f : (a,b) \to (0,\infty)$ is bounded, then $(\overline{M},\overline{g})$ is past and future null geodesically incomplete. Thus, assuming that a and b are finite, one-parameter families $(\overline{M},\overline{g}(s)) = (\overline{M}, -dt^2 \oplus f(s)h)$ of past and future null geodesically incomplete space–times may easily be constructed. Choosing the one-parameter family of functions $f(s) : (a,b) \to (0,\infty)$ suitably, the curve

$s \to g(s) = -dt^2 \oplus f(s)h$ of metrics will *not* be a continuous curve in $\mathrm{Lor}(\overline{M})$ in the fine C^r topologies. Thus the space–times $(\overline{M}, \overline{g}(0))$ and $(\overline{M}, \overline{g}(s))$ may be far apart in $\mathrm{Lor}(\overline{M})$ for $s \neq 0$.

Notice that if the Riemannian manifold (H, h) is geodesically incomplete, then $\overline{M} = (a, b) \times_f H$ may be null geodesically incomplete even if both integrals in Theorem 3.70 diverge. On the other hand, if the completeness of (H, h) is assumed, the following necessary and sufficient condition for the null geodesic incompleteness of $\overline{M} = (a, b) \times_f H$ may be obtained from the proof of Theorem 3.70.

Remark 3.71. Let $\overline{M} = (a, b) \times_f H$ be a Lorentzian warped product with Lorentzian metric $\overline{g} = -dt^2 \oplus fh$, where (H, h) is a complete Riemannian manifold and $-\infty \leq a < b \leq +\infty$. Let $S(t) = \sqrt{f(t)}$ as above. Then $(\overline{M}, \overline{g})$ is past (respectively, future) null geodesically incomplete if and only if $\lim\limits_{t \to a^+} \int_t^{\omega_0} S(s)\, ds$ is finite (respectively, $\lim\limits_{t \to b^-} \int_{\omega_0}^t S(s)\, ds$ is finite).

In Powell (1982), a more comprehensive study is made of nonspacelike geodesic completeness of Lorentzian warped products, beginning with the observation that a geodesic in $M \times_f H$ projects to a pregeodesic in (H, h) and continuing with the observation that if $\widetilde{\gamma}$ and $\widetilde{\beta}$ are unit speed geodesics in (H, h) and $(\gamma, \widetilde{\gamma})$ is a pregeodesic of $M \times_f H$, then $(\gamma, \widetilde{\beta})$ is also a pregeodesic of $M \times_f H$. Then Powell shows that if the Lorentzian warped product $M \times_f H$ is timelike (respectively, null or spacelike) geodesically complete, then (H, h) is Riemannian complete and further (M, g) is timelike (respectively, null or spacelike) complete.

A second aspect of Powell (1982) is the study of timelike geodesic completeness for warped products of the form $\overline{M} = (a, b) \times_f H$ with metrics $\overline{g} = -dt^2 \oplus fh$, corresponding to Theorem 3.70 and Remark 3.71 above for null completeness. As remarked above, the completeness of the timelike geodesics of the form $\overline{\gamma}(t) = (t, y_0)$ for some $y_0 \in H$, which Powell terms "stationary," is entirely dependent on whether $a = -\infty$ or a is finite, and/or $b = +\infty$ or b is finite. Also, if a and b are both finite, then all timelike geodesics are both past and future incomplete independent of choice of warping function. However, for a suitable warping function, even if $a = -\infty$ and $b = +\infty$, it can be arranged for all non-stationary timelike geodesics to be incomplete. More precisely, Powell shows that if (H, h) is complete and $\omega_0 \in (a, b)$, then all future directed non-stationary timelike geodesics are future (respectively, past) complete if and only if

$$\int_{\omega_0}^{b} \left(\frac{f(t)}{1 + f(t)} \right)^{\frac{1}{2}} dt = +\infty, \qquad \text{respectively,} \quad \int_{a}^{\omega_0} \left(\frac{f(t)}{1 + f(t)} \right)^{\frac{1}{2}} dt = +\infty.$$

From this result and our above results on null completeness, relationships may be derived between null and timelike completeness for this class of warped products. (In general, these types of geodesic completeness are logically independent (cf. Theorem 6.4). In particular, null completeness of $\mathbb{R} \times_f H$ does not imply timelike completeness.) However, if the warping function f for $\overline{M} = \mathbb{R} \times_f H$ is bounded from above and (H, h) is Riemannian complete, then the timelike and null completeness are equivalent.

In singularity theory in general relativity, conditions on the curvature tensor of $(\overline{M}, \overline{g})$ which are discussed in Chapters 2 and 12, called the *generic condition* and the *timelike convergence condition*, are considered. These two conditions guarantee that if a nonspacelike geodesic γ may be extended to be defined for all positive and negative values of an affine parameter and $\dim \overline{M} \geq 3$, then γ contains a pair of conjugate points. Hence these curvature conditions may be combined with geometric or physical assumptions, such as $(\overline{M}, \overline{g})$ is causally disconnected or $(\overline{M}, \overline{g})$ contains a closed trapped set, to show that $(\overline{M}, \overline{g})$ is nonspacelike geodesically incomplete (cf. Section 12.4). Since $(\overline{M}, \overline{g})$ satisfies the generic condition and strong energy condition if all nonspacelike Ricci curvatures are positive, it is thus of interest to consider conditions on the warping function f of a Lorentzian warped product which guarantee that $(\overline{M}, \overline{g})$ has everywhere positive nonspacelike Ricci curvatures. The assumption $\dim \overline{M} \geq 3$ made in singularity theory is necessary for null conjugate points to exist since no null geodesic in any two-dimensional Lorentzian manifold contains a pair of conjugate points.

We now give the formulas for the curvature tensor R and Ricci curvature tensor Ric for the warped product space–time $(M \times_f H, \overline{g})$ where $\overline{g} = g \oplus fh$.

As above, let ∇^1 [respectively, ∇^2] denote the covariant derivative of (M, g) [respectively, (H, h)]. Also let $\phi = \ln f$ and recall that $\operatorname{grad}\phi$ denotes the gradient of ϕ on (M, g). As before, we will decompose tangent vectors x in $T_{\overline{p}}(M \times H)$ as $x = (x_1, x_2)$. Let R^1 [respectively, R^2] denote the curvature tensor of (M, g) [respectively, (H, h)]. Given tangent vectors $x_1, y_1 \in T_pM$, define the Hessian tensors H_ϕ and h_ϕ by

$$H_\phi(x_1) = \nabla^1_{x_1} \operatorname{grad}\phi \tag{3.19}$$

and

$$h_\phi(x_1, y_1) = g(\nabla^1_{x_1} \operatorname{grad}\phi, y_1). \tag{3.20}$$

We will also write $\|\operatorname{grad}\phi\|^2 = g(\operatorname{grad}\phi, \operatorname{grad}\phi)$. Using the sign convention

$$R(X, Y)Z = \overline{\nabla}_X\overline{\nabla}_Y Z - \overline{\nabla}_Y\overline{\nabla}_X Z - \overline{\nabla}_{[X,Y]} Z$$

for the curvature tensor and substituting from formula (3.17), one obtains the formula

$$\begin{aligned} R(x, y)z = {} & R^1(x_1, y_1)z_1 + R^2(x_2, y_2)z_2 + \frac{1}{2}\,[h_\phi(x_1, z_1)y_2 \\ & - h_\phi(y_1, z_1)x_2 + \overline{g}(x_2, y_2)H_\phi(y_1) - \overline{g}(y_2, z_2)H_\phi(x_1)] \\ & + \frac{1}{4}\left\{\left[x_1(\phi)z_1(\phi) + \overline{g}(x_2, z_2)\|\operatorname{grad}\phi\|^2(p)\right] y_2\right. \\ & - \left[y_1(\phi)z_1(\phi) + \overline{g}(y_2, z_2)\|\operatorname{grad}\phi\|^2(p)\right] x_2 \\ & \left. + \left[y_1(\phi)\overline{g}(x_2, z_2) - x_1(\phi)\overline{g}(y_2, z_2)\right] \operatorname{grad}\phi(p)\right\} \end{aligned} \tag{3.21}$$

where $x, y, z \in T_{(p,q)}(M \times H)$.

Suppose now that $\dim M = m$ and $\dim H = n$. To calculate the Ricci curvature at $\overline{p} = (p, q) \in M \times H$, let $\{e_1, e_2, \ldots, e_m\}$ be a basis for T_pM with $g(e_1, e_1) = -1$, $g(e_j, e_j) = 1$ for $2 \leq j \leq m$, and $g(e_i, e_j) = 0$ if $i \neq j$. Also, let $\{e_{m+1}, \ldots, e_{n+m}\}$ be a $\overline{g}$-orthonormal basis for T_qH. Then for any $x, y \in T_{\overline{p}}(M \times H)$, we have

$$\operatorname{Ric}(x, y) = -\overline{g}\,(R(e_1, x)y, e_1) + \sum_{j=2}^{n+m} \overline{g}(R(e_j, x)y, e_j).$$

The d'Alembertian $\Box\phi$ of ϕ may also be calculated as

$$\Box\phi(p) = -h_\phi(e_1, e_1) + \sum_{j=2}^{m} h_\phi(e_j, e_j).$$

Using (3.21), it then follows that

$$\begin{aligned} \operatorname{Ric}(x, y) &= \operatorname{Ric}^1(x_1, y_1) + \operatorname{Ric}^2(x_2, y_2) \\ &\quad - \overline{g}(x_2, y_2)\left[\frac{1}{2}\Box\phi(p) + \frac{\dim H}{4}\|\operatorname{grad}\phi(p)\|^2\right] \\ &\quad - \frac{\dim H}{2} h_\phi(x_1, y_1) - \frac{\dim H}{4} x_1(\phi) y_1(\phi) \end{aligned} \tag{3.22}$$

where $x = (x_1, y_1)$, $y = (y_1, y_2) \in T_{(p,q)}(M \times H)$, and Ric^1 and Ric^2 denote the Ricci curvature tensors of (M, g) and (H, h), respectively.

We now restrict to the case $\overline{M} = (a, b) \times_f H$ with warped product metric $\overline{g} = -dt^2 \oplus fh$. In this case, $\Box\phi(t) = -\phi''(t)$ and $\|\operatorname{grad}\phi(t)\|^2 = -[\phi'(t)]^2$. Thus we obtain from (3.22) for $\overline{v} = (0, v) \in T_{(t,q)}((a, b) \times H)$ that

$$\operatorname{Ric}(\overline{v}, \overline{v}) = \operatorname{Ric}^2(v, v) + \overline{g}(v, v)\left\{\frac{1}{2}\phi''(t) + \frac{\dim H}{4}[\phi'(t)]^2\right\}. \tag{3.23}$$

If $x = \partial/\partial t|_t + v \in T_{(t,q)}((a, b) \times H)$ with $v \in T_q H$, we obtain

$$\begin{aligned} \operatorname{Ric}(x, x) &= \operatorname{Ric}^2(v, v) + \overline{g}(v, v)\left\{\frac{1}{2}\phi''(t) + \frac{\dim H}{4}[\phi'(t)]^2\right\} \\ &\quad + \left\{-\frac{\dim H}{2}\phi''(t) - \frac{\dim H}{4}[\phi'(t)]^2\right\}. \end{aligned} \tag{3.24}$$

Both bracketed terms in formulas (3.23) and (3.24) will be positive provided that

$$-[\phi'(t)]^2 \dim H < 2\phi''(t) < -[\phi'(t)]^2 \tag{3.25}$$

for all $t \in (a, b)$. Thus if $\operatorname{Ric}^2(v, v) \geq 0$ for all $v \in TH$ and condition (3.25) holds, the space–time $(\overline{M}, \overline{g})$ will have everywhere positive Ricci curvatures. A globally hyperbolic family of such space–times is provided by warped products $\overline{M} = (0, \infty) \times_f H$, where (H, h) is a complete Riemannian manifold of nonnegative Ricci curvature and $\overline{g} = -dt^2 \oplus fh$ with $f(t) = t^r$ for a fixed constant $r \in \mathbb{R}$ satisfying $(2/\dim H) < r < 2$. If (H, h) is taken to be $\mathbb{R}^3$ with the usual Euclidean metric and $r = 4/3$, we recover the Einstein–de Sitter universe of cosmology theory [cf. Hawking and Ellis (1973, p. 138), Sachs and Wu (1977a, Proposition 6.2.7 ff.)].

We may also obtain the following condition on $\phi = \ln f$ for positive nonspacelike Ricci curvature if the Ricci tensor of (H, h) is bounded from below.

Proposition 3.72. *Let $\overline{M} = (a, b) \times_f H$ with $n = \dim H \geq 2$, $\overline{g} = -dt^2 \oplus fh$, and $\phi = \ln f$. Suppose that there exists some constant $\lambda \in \mathbb{R}$ such that $\mathrm{Ric}^2(v, v) \geq \lambda h(v, v)$ for all $v \in TH$. Then if*

$$2\phi''(t) < \min\left\{-(\phi'(t))^2,\ 4(n-1)^{-1}\lambda e^{-\phi(t)}\right\} \tag{3.26}$$

for all $t \in (a, b)$, the Lorentzian warped product $(\overline{M}, \overline{g})$ has everywhere positive nonspacelike Ricci curvature.

Proof. It suffices to show that $\mathrm{Ric}(x, x) > 0$ for all nonspacelike tangent vectors x of the form $x = \partial/\partial t + v \in T(M \times H)$, $v \in TH$. Since $\overline{g}(x, x) \leq 0$ and $\overline{g}(\partial/\partial t, \partial/\partial t) = -1$, we have $\beta = \overline{g}(v, v) \leq 1$. Hence $0 \leq \beta \leq 1$. Then $h(v, v) = \beta e^{-\phi}$, and we obtain from (3.24) that

$$\mathrm{Ric}(x, x) \geq \beta e^{-\phi}\lambda + \left[\frac{\beta}{2} - \frac{n}{2}\right]\phi'' + \frac{n}{4}(\beta - 1)(\phi')^2. \tag{3.27}$$

Thus $\mathrm{Ric}(x, x) > 0$ provided $\phi'' < G(\beta)$ for all $\beta \in [0, 1]$, where

$$G(\beta) = \frac{4\beta e^{-\phi}\lambda - n(1-\beta)(\phi')^2}{2(n-\beta)}.$$

Calculating $G''(\beta)$, one finds that $G'(\beta)$ does not change sign in $[0, 1]$. Thus $G(\beta)$ obtains its minimum on $[0, 1]$ for $\beta = 0$ or $\beta = 1$. Hence $\mathrm{Ric}(x, x) > 0$ provided that $\phi'' < \min\{G(0), G(\beta)\}$, which yields inequality (3.26). $\square$

We now consider the scalar curvature of warped product manifolds of the form $\overline{M} = \mathbb{R} \times_f H$, $\overline{g} = -dt^2 \oplus fh$. We will let $n = \dim H$ below. Given $(t, p) \in \overline{M}$, choose $e_j \in T_pH$ for $1 \leq j \leq n$ such that if $\overline{e}_j = (0, e_j) \in T_{(t,p)}\overline{M}$, then $\{\overline{\partial/\partial t} = (\partial/\partial t, 0_p), \overline{e}_1, \ldots, \overline{e}_n\}$ forms a $\overline{g}$-orthonormal basis for $T_{(t,p)}\overline{M}$.

Hence $\left\{\sqrt{f(t)}e_1, \ldots, \sqrt{f(t)}e_n\right\}$ forms an h-orthonormal basis for T_pH. Thus if $\tau : \overline{M} \to \mathbb{R}$ and $\tau_H : H \to \mathbb{R}$ denote the scalar curvature functions of $(\overline{M}, \overline{g})$ and (H, h) respectively, we have

$$\tau(t,p) = -\operatorname{Ric}\left(\frac{\overline{\partial}}{\partial t}, \frac{\overline{\partial}}{\partial t}\right) + \sum_{j=1}^{n} \operatorname{Ric}(\overline{e}_j, \overline{e}_j)$$

and

$$\tau_H(p) = f(t) \sum_{j=1}^{n} \operatorname{Ric}^2(e_j, e_j).$$

Now formulas (3.23) and (3.24) above simplify to

$$\operatorname{Ric}\left(\frac{\overline{\partial}}{\partial t}, \frac{\overline{\partial}}{\partial t}\right) = -\frac{n}{2}\phi''(t) - \frac{n}{4}[\phi'(t)]^2 \tag{3.28}$$

and

$$\operatorname{Ric}(\overline{e}_j, \overline{e}_j) = \operatorname{Ric}^2(e_j, e_j) + \frac{1}{2}\phi''(t) + \frac{n}{4}[\phi'(t)]^2 \tag{3.29}$$

for $1 \leq j \leq n$. Consequently, we obtain the formula

$$\tau(t,p) = \frac{1}{f(t)}\tau_H(p) + n\phi''(t) + \frac{1}{4}(n^2 + n)[\phi'(t)]^2.$$

Recalling that $\phi(t) = \ln f(t)$, this may be rewritten as

$$\tau(t,p) = \frac{1}{f(t)}\tau_H(p) + n\frac{f''(t)}{f(t)} + \frac{1}{4}(n^2 - 3n)\left[\frac{f'(t)}{f(t)}\right]^2 \tag{3.30}$$

where $\dim H = n$ as above. In particular, in the case that $n = 3$ as in general relativity, we obtain the simpler formula

$$\tau(t,p) = \frac{1}{f(t)}\tau_H(p) + 3\frac{f''(t)}{f(t)}. \tag{3.31}$$

Example 3.73. With the formulas of this section in hand, we are now ready to give an example of a 1-parameter family $\overline{g}_\lambda$ of nonisometric Einstein metrics for $\mathbb{R}^{n+1}$ such that for $\lambda = 0$, $(\mathbb{R}^{n+1}, g_0)$ is Minkowski space–time of dimension $n + 1$. Let $(\mathbb{R}^n, h)$ be Euclidean n-space with the usual Euclidean metric $h = dx_1{}^2 + dx_2{}^2 + \cdots + dx_n{}^2$, and put $M_\lambda = \mathbb{R}^{n+1} = \mathbb{R} \times_f \mathbb{R}^n$ with the

Lorentzian metric $\overline{g}_\lambda = -dt^2 \oplus e^{\lambda t}h$, i.e., $f(t) = e^{\lambda t}$. By Theorem 3.70, for all $\lambda > 0$ the space–time $(\mathbb{R}^{n+1}, \overline{g}_\lambda)$ is future null geodesically complete but past null geodesically incomplete, and for all $\lambda < 0$, the space–time $(\mathbb{R}^{n+1}, \overline{g}_\lambda)$ is past null geodesically complete but future null geodesically incomplete. Using formulas (3.28), (3.29), and (3.30), we obtain

$$\operatorname{Ric}(\overline{g}_\lambda) = \frac{n\lambda^2}{4}\,\overline{g}_\lambda \tag{3.32}$$

and

$$\tau_{\overline{g}_\lambda} = \frac{1}{4}(n^2 + n)\lambda^2. \tag{3.33}$$

Thus if $\lambda \neq 0$, $(\overline{M}_\lambda, \overline{g}_\lambda)$ is an Einstein space–time with constant positive scalar curvature.

Example 3.74. Let $\overline{M}_\lambda = (0, \infty) \times_f \mathbb{R}^3$, where $\overline{g}_\lambda = -dt^2 \oplus fh$ with $f(t) = \lambda t$, $\lambda > 0$, and h the usual Euclidean metric on $\mathbb{R}^3$. It is then immediate from formula (3.31) that $\tau(g_\lambda) = 0$ for all $\lambda > 0$. Since $\phi(t) = \ln(\lambda t)$, it may be checked using formulas (3.28) and (3.29) that $(\overline{M}_\lambda, \overline{g}_\lambda)$ is neither Ricci flat nor Einstein for any $\lambda > 0$. Also we have for any $\lambda > 0$ that

$$\operatorname{Ric}\left(\frac{\overline{\partial}}{\partial t}, \frac{\overline{\partial}}{\partial t}\right) = \frac{3}{4}t^{-2}$$

for all $t > 0$. It follows that the space–times $(\overline{M}_\lambda, \overline{g}_\lambda)$ are "inextendible across" $\{0\} \times \mathbb{R}^3$ (cf. Section 6.5). Also, $(\overline{M}_\lambda, \overline{g}_\lambda)$ is future null geodesically complete by Theorem 3.70.

3.7 Semi-Riemannian Local Warped Product Splittings

In each of the exact solutions to Einstein's equations which are presented as warped product manifolds, the warped product decomposition emerges as a natural mathematical expression of assumed physical symmetries. Moreover, formulas for warped product curvatures (cf. Proposition 3.76) indicate that any semi-Riemannian manifold (M, g) must possess certain measures of symmetry and flatness in order to be (locally or globally) isometric to a warped product

$B \times_f F$. In this section we identify geometric conditions on a semi-Riemannian manifold (M, g) which are necessary and sufficient to ensure that (M, g) is locally isometric to a warped product $B \times_f F$. We shall call such a local isometry a "local warped product splitting."

There are a number of different "splitting" or decomposition theorems throughout differential geometry. For example, the splitting theorem of Geroch (1970a), Theorem 3.17 in this chapter, demonstrates that a globally hyperbolic space–time may be written as a particular type of *topological* product but not necessarily as a *metric* product. By contrast, a well-known result of de Rham [cf. Kobayashi–Nomizu (1963, p. 187)] asserts that a complete simply connected Riemannian manifold which has a reducible holonomy representation is isometric to a Riemannian product. Along very different lines, Chapter 14 provides the following Lorentzian analogue of the Cheeger–Gromoll Splitting Theorem: if (M, g) is a space–time of dimension $n \geq 3$ which (1) is globally hyperbolic or timelike geodesically complete, (2) satisfies the timelike convergence condition, and (3) contains a complete timelike line, then (M, g) is isometric to a product $(\mathbb{R} \times V, -dt^2 \oplus h)$, where (V, h) is a complete Riemannian manifold. Since a product manifold is trivially a warped product, either of these last two results clearly provides sufficient conditions to ensure that a manifold is globally isometric to a warped product. However, the examples $(S^1 \times_f H, \overline{g})$ of non-globally hyperbolic warped product space–times discussed in Section 3.6 indicate that causal assumptions such as global hyperbolicity are not necessary for the existence of a global warped product splitting.

The global splitting question typically involves rather delicate topological considerations; our focus in this section will be on the simpler *local warped product splitting question: given a semi-Riemannian manifold* (M, g) *and a point* $p \in M$*, what conditions are necessary and sufficient for the existence of an open neighborhood* $\mathcal{U}$ *of* p *such that the submanifold* $(\mathcal{U}, g|_{\mathcal{U}})$ *is isometric to a warped product* $B \times_f F$? The following assumption will be needed.

Convention 3.75. It will be assumed throughout the remainder of this section that the neighborhood $\mathcal{U}$ mentioned above is a connected, simply connected, open set.

The geometry of a warped product $B \times_f F$ is expressed through the geometries of the base (B, g_B) and fiber (F, g_F) and various derivatives and integrals of the warping function $f \in \mathfrak{F}(B)$. We will consider warped products $M = B \times_f F$ where *both* (B, g_B) and (F, g_F) may be semi-Riemannian manifolds (thus generalizing the *Lorentzian* warped products of Definition 3.51). The symbols $\pi : M \to B$ and $\sigma : M \to F$ denote the standard projections. As in portions of Section 3.6, we will find it convenient to consider the square root S of the warping function f. *Throughout this section, we will adhere to the convention that* $S(b) = \sqrt{f(b)}$ *for* $b \in B$ *where* $M = B \times_f F$ *denotes a warped product with metric tensor*

$$g = g_B + f\, g_F = g_B + S^2\, g_F.$$

The function S will be called the *root warping function.*

As defined in Section 3.6, the *lift* of $f \in \mathfrak{F}(B)$ to a function $\widetilde{f} \in \mathfrak{F}(M)$ is defined by the formula $\widetilde{f} = f \circ \pi$. The lifted function will simply be denoted by f as well, when no ambiguity results. A vector field $V \in \mathfrak{X}(B)$ is lifted to M by defining $\widetilde{V} \in \mathfrak{X}(M)$ in such a manner that at each $(p, q) \in M$, $\widetilde{V}(p, q)$ is the unique vector in the tangent space $T_{(p,q)}M$ such that both $d\pi(\widetilde{V}) = V$ and $d\sigma(\widetilde{V}) = 0$. Similar definitions apply for lifts from F to M. We shall also denote lifted vector fields without the tildes, and we write $V \in \mathfrak{L}(F)$ to denote a vector field on M lifted from F. More generally, vectors tangent to *leaves* $B \times q$ are called *horizontal* while those tangent to *fibers* $p \times F$ are called *vertical.* Lifts of covariant tensors on B and F are now defined in the obvious way through the use of the pullbacks π^* and σ^*.

Let D denote the Levi–Civita connection on M, and use ∇ to denote the Levi-Civita connections on both B and F. The curvature tensors on $B \times_f F = M$ may be characterized through their actions on lifted horizontal and vertical vector fields. In the following proposition, the symbols ${}^B R$ and ${}^F R$ denote the lifts to M of the Riemannian curvature tensors on B and F, respectively. The symbol H^S will be used to denote $\widetilde{H^S}$, the lift to M of the Hessian of S. Note that in general $\widetilde{H^S}(X, Y) = H^{\widetilde{S}}(X, Y)$ only for *horizontal* vector fields X, Y. For notational simplicity, the bracket notation $\langle \quad , \quad \rangle$ will be occasionally used to denote the metric g on M; the metrics on B and F will always be denoted by g_B and g_F.

Some basic curvature formulas for warped product manifolds are now given in Proposition 3.76 for ease of reference. The standard reference for this material is O'Neill (1983).

Proposition 3.76. *Let the semi-Riemannian warped product $M = B \times_f F$ have Riemannian curvature tensor R, Ricci curvature* Ric, *and root warping function $S = \sqrt{f}$. Assume $X, Y, Z \in \mathfrak{L}(B)$ and $U, V, W \in \mathfrak{L}(F)$.*

(1) *If $h \in \mathfrak{F}(B)$, then the gradient of the lift $h \circ \pi$ of h to $M = B \times_f F$ is the lift to M of the gradient of h on B, i.e.,* $\operatorname{grad}(\widetilde{h}) = \widetilde{\operatorname{grad} h}$.

(2) *$D_X Y \in \mathcal{L}(B)$ is the lift of $\nabla_X Y$ on B, i.e., $D_{\widetilde{X}} \widetilde{Y} = (\widetilde{\nabla_X Y})$.*

(3) *$D_X V = D_V X = \left(\frac{XS}{S}\right) V$.*

(4) $\operatorname{Ric}(X, Y) = {}^B\operatorname{Ric}(X, Y) - \left(\frac{d}{S}\right) H^S(X, Y)$ *where $d = \dim F$.*

(5) $\operatorname{Ric}(V, X) = 0$.

(6) $\operatorname{Ric}(V, W) = {}^F\operatorname{Ric}(V, W) - \langle V, W\rangle S^\sharp$
where $S^\sharp = \frac{\Delta S}{S} + (d-1)\frac{\langle \operatorname{grad} S, \operatorname{grad} S\rangle}{S^2}$, $d = \dim F$, and ΔS is the Laplacian of the root warping function S on B.

Consider first the local warped product splitting question in dimension two. Assume a semi-Riemannian surface (M, g) is a warped product so that in the appropriate local coordinates (y^1, y^2) adapted to B and F, the metric is given by

$$g = \epsilon_1 dy^1 \otimes dy^1 + \epsilon_2 [S(y^1)]^2 dy^2 \otimes dy^2 \tag{3.34}$$

where $\epsilon_i = \pm 1$, $i = 1, 2$. It is immediate that $\partial/\partial y^2$ is a local Killing field (recall the elementary fact that a coordinate vector field $\partial/\partial x^k$ is Killing if and only if $\frac{\partial g_{ij}}{\partial x^k} = 0$ for all i, j). Each $(\partial/\partial y^1)$–integral curve is a geodesic of M (more generally, each leaf $B \times q$ of a warped product is a totally geodesic submanifold). Thus $\partial/\partial y^2$ restricted to γ is a Jacobi field for each such coordinate geodesic γ, and of course, $\langle \partial/\partial y^1, \partial/\partial y^2\rangle = 0$ along γ. Further, direct calculation shows that $\partial/\partial y^1$ is *irrotational*—that is, $\operatorname{curl}(\partial/\partial y^1) = 0$, where the curl of a vector field is the skew-symmetric (0,2) tensor defined through the formula

$$[\operatorname{curl} X](Y, Z) = \langle D_Y X, Z\rangle - \langle Y, D_Z X\rangle \tag{3.35}$$

for all vector fields X, Y, and Z. The unit vector field $\frac{1}{S}(\partial/\partial y^2)$ does not have vanishing curl, in general. These observations lead to the following result.

Lemma 3.77. *Given a two-dimensional semi-Riemannian manifold* (M, g) *and a point* $p \in M$, *p has a neighborhood* $\mathcal{U}$ *such that* $(\mathcal{U}, g|_{\mathcal{U}})$ *is isometric to a warped product if and only if there exists a non-vanishing Killing field on an open neighborhood of* p.

Proof. The local existence of a nonvanishing Killing field about any point of a two-dimensional semi-Riemannian warped product was noted above.

Conversely, suppose there exists a neighborhood $\mathcal{U}$ about $p \in M$ having a nonvanishing Killing field V. If the Killing field is null, then it is well known that $(\mathcal{U}, g|_{\mathcal{U}})$ is isometric to (a portion of) Minkowski two-space $\mathbb{R}^2_1$ and hence is (trivially) a warped product.

If the Killing field V is not null, then we may complete the classical construction of local geodesic (or Fermi) coordinates (x^1, x^2) [cf. do Carmo (1976)] such that $V = \partial/\partial x^2$ on $\mathcal{U}$ and x^1 measures g-arc length along the geodesics orthogonal to the integral curves of V. In these coordinates the metric assumes the form

$$ds^2 = \epsilon_1\, dx^1 \otimes dx^1 + \epsilon_2\, F(x^1, x^2) dx^2 \otimes dx^2$$

where $\epsilon_i = \pm 1$, $i = 1, 2$. Since $V = \partial/\partial x^2$ is Killing, we must have $\frac{\partial F}{\partial x^2} = 0$, yielding

$$ds^2 = \epsilon_1\, dx^1 \otimes dx^1 + \epsilon_2\, F(x^1) dx^2 \otimes dx^2$$

and leading to the desired local warped product representation. □

In generalizing the preceding result to higher dimensions, the existence of a Killing field V must be supplemented by an integrability condition which allows the construction of leaves $B \times q$ orthogonal to V. This integrability condition holds trivially in dimension two but need not hold in higher dimensions where, in general, a Killing field need not be irrotational. It is also necessary to include the additional assumption that the Killing field be nonnull.

Recall that a space–time (M, g) is called *static* if there exists on M a nowhere zero timelike Killing field X such that the distribution of $(n-1)$-planes orthogonal to X is integrable. The following result parallels the formal construction of static space–times as presented in a number of texts on general relativity [cf. Sachs and Wu (1977a)].

Lemma 3.78. *Let (M, g) be an n-dimensional semi-Riemannian manifold with $n \geq 3$, and let $m \in M$ be any point. There exists a local isometry of a neighborhood $\mathcal{U}$ of m with a warped product $B \times_f F$ having $\dim F = 1$ and $\dim B = (n-1)$ if and only if*

(1) *There exists a nonnull non-vanishing Killing field V on a neighborhood of m such that*

(2) *the unit vector field $U = \frac{V}{\|V\|}$ satisfies $[\operatorname{curl} U](X, Y) = 0$ for all $X, Y \perp U$ on this neighborhood of m.*

Proof. Assume M is locally isometric to a warped product $B \times {}_f F$ with $\dim F = 1$. The point m may then be uniquely written as $m = (b, q)$ with $b \in B$ and $q \in F$. Let $B \times q$ and $p \times F$ denote the leaf and fiber through m, respectively. In local coordinates adapted to the submanifolds $B \times q$ and $p \times F$ such that $g_F = \epsilon_n dx^n \otimes dx^n$ (i.e., x^n is arc length on F), we have

$$g = \pi^* g_B + \epsilon_n S^2 dx^n \otimes dx^n, \qquad S \in \mathfrak{F}(B),$$

where $\epsilon_n = \operatorname{sign}(g_F(\partial/\partial x^n, \partial/\partial x^n))$. Now, $\partial/\partial x^n$ is a nonnull non-vanishing Killing field on the domain $\mathcal{U}$ of the chart. Furthermore, $\frac{1}{S}(\partial/\partial x^n)$ is a unit magnitude vector field, and for coordinate vector fields $\partial/\partial x^i$, $\partial/\partial x^j$ orthogonal to $\partial/\partial x^n$, $1 \leq i, j \leq n-1$,

$$\begin{aligned}\left[\operatorname{curl}\left(\frac{1}{S}\frac{\partial}{\partial x^n}\right)\right]\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) &= \left\langle D_{\partial/\partial x^i}\left(\frac{1}{S}\frac{\partial}{\partial x^n}\right), \frac{\partial}{\partial x^j}\right\rangle \\ &\quad - \left\langle D_{\partial/\partial x^j}\left(\frac{1}{S}\frac{\partial}{\partial x^n}\right), \frac{\partial}{\partial x^i}\right\rangle \\ &= 0,\end{aligned}$$

where formula (3) of Proposition 3.76 was used to evaluate the covariant derivatives. It follows that for any *horizontal* vector fields $X = \sum_{i=1}^{n-1} X^i(\partial/\partial x^i)$ and $Y = \sum_{i=1}^{n-1} Y^i(\partial/\partial x^i)$ orthogonal to U, $\left[\operatorname{curl}\left(\frac{1}{S}\frac{\partial}{\partial x^m}\right)\right](X, Y) = 0$ even if the fields are not lifts from B, i.e., even if some coefficient functions X^i and Y^j

are functions of x^n as well as $x^1, x^2, \ldots, x^{n-1}$. Thus, conditions (1) and (2) are clearly necessary.

For sufficiency, assume a nonnull non-vanishing Killing field V exists on an open neighborhood $\mathcal{U} \subseteq M$ of the point m. At each $p \in \mathcal{U}$ choose $(V_p)^\perp \subseteq T_p\mathcal{U}$, giving an $(n-1)$-dimensional distribution $\mathcal{D}$ on U. Condition (2) ensures that $\mathcal{D}$ is involutive; applying Frobenius' Theorem, we may integrate this distribution to provide, through each point of $\mathcal{U}$, a local integral $(n-1)$-submanifold of M having $\mathcal{D}$ as its tangent space. Denote by B the integral $(n-1)$-submanifold through the distinguished point m having $\mathcal{D}(m)$ as its tangent space. Denote by F the unique integral curve for U passing through m.

Clearly, $\mathcal{U}$ may be thought of as a topological product $B \times F$ by shrinking $\mathcal{U}$ if necessary. We must show the induced metric $g\,|\,\mathcal{U}$ can be expressed as a warped product of appropriate metrics on B and F. Let $\xi = (x^1, x^2, \ldots, x^n)$ be adapted coordinates about m, so that $B = \{(x^1, x^2, \ldots, x^n)\,|\,x^n = 0\}$, $U = \partial/\partial x^n$, and m has coordinates $(0, 0, \ldots, 0)$. We shrink $\mathcal{U}$ if necessary so that $\operatorname{dom}(\xi) = \mathcal{U}$.

A point in $\mathcal{U}$ may be uniquely specified as an ordered pair (p, q) with $p \in B$ and $q \in F$. For each fixed $p \in B$, $\|V\,|_{(p,q)}\| = |g(V|_{(p,q)}, V|_{(p,q)})|^{\frac{1}{2}}$ has constant value as q varies over F since $V = \partial/\partial x^n$ is Killing. Thus, we may unambiguously define a positive function $S \in \mathfrak{F}(B)$ by

$$\begin{aligned} S(p) &= |g(V\,|_{(p,q)}, V\,|_{(p,q)})|^{\frac{1}{2}} \\ &= \left| g\left(\left.\frac{\partial}{\partial x^n}\right|_{(p,q)}, \left.\frac{\partial}{\partial x^n}\right|_{(p,q)} \right) \right|^{\frac{1}{2}} \\ &= |g_{nn}(p,q)|^{\frac{1}{2}}, \end{aligned}$$

which is independent of q for each fixed p. We define a metric g_B on B by restriction of the metric on M: $g_B = g\,|_B$. The metric on F is defined as $g_F = \epsilon_n dx^n \otimes dx^n$, where $\epsilon_n = \operatorname{sign} g(V, V)$.

Now $\partial/\partial x^n \perp \partial/\partial x^i$ for $i = 1, 2, \ldots, (n-1)$ by construction. Furthermore, the coordinate patch ξ provides us with natural projections π and σ from $\mathcal{U}$ into B and F, respectively: $\pi : (x^1, x^2, \ldots, x^n) \to (x^1, x^2, \ldots, x^{n-1}, 0)$, and $\sigma : (x^1, x^2, \ldots, x^n) \to (0, 0, \ldots, 0, x^n)$. For arbitrary $X, Y \in \mathfrak{X}(\mathcal{U})$ with coordinate basis expansions $X = \sum X^i(\partial/\partial x^i)$, $Y = \sum Y^i(\partial/\partial x^i)$, we have

$$\begin{aligned} g(X,Y) &= g\left(\sum_{i=1}^{n} X^i(\partial/\partial x^i), \sum_{j=1}^{n} Y^j(\partial/\partial x^j)\right) \\ &= g\left(\sum_{i=1}^{n-1} X^i(\partial/\partial x^i), \sum_{j=1}^{n-1} Y^j(\partial/\partial x^j)\right) + g(X^n\partial/\partial x^n, Y^n\partial/\partial x^n) \\ &= g_B(\pi_*X, \pi_*Y) + g_{nn}X^nY^n \\ &= g_B(\pi_*X, \pi_*Y) + [S(p)]^2 g_F(\sigma_*X, \sigma_*Y). \quad \square \end{aligned}$$

The "dimensional dual" of the preceding result asks for conditions necessary and sufficient to ensure that (M, g) is locally isometric to a warped product $B \times_f F$ with $\dim B = 1$. One answer to this question involves a construction which bears strong similarities to the formal development of the Robertson–Walker cosmological models [cf. O'Neill (1983)]. Recall that a signal feature of Robertson–Walker space–time is the presence of a proper time synchronizable geodesic observer field U. Since the observer field U is irrotational, the *infinitesimal* rest spaces of U may be integrated to provide *local* rest spaces. Through a construction quite similar to that of the preceding lemma, it is possible to verify the following [cf. Easley (1991)]: given an n-dimensional ($n \geq 3$) semi-Riemannian manifold and point $p \in M$, the point p has a neighborhood $\mathcal{U} \subseteq M$ such that $(\mathcal{U}, g|_{\mathcal{U}})$ is isometric to a warped product $B \times_f F$ with $\dim B = 1$ and $\dim F = (n-1)$ if and only if (1) there exists an irrotational unit vector field U (either spacelike or timelike) on a neighborhood of p such that (2) the flow Ψ induced by U acts as a positive homothety on the local $(n-1)$-dimensional submanifolds which are everywhere orthogonal to U near p. Condition (2) is of course equivalent to a number of different geometric conditions expressible in terms of curvature and the connection.

The local splitting problem in the context of Ricci flat (M, g) has an extremely restricted class of solutions. We consider first the only case in dimension four which will not be subsumed under more general results, namely, the case when M is locally isometric to a warped product with base and fiber each of dimension two. The following computational lemma is a necessary preliminary step. In the following result, the symbols ${}^FK$ and ${}^BK$ will be used to denote the sectional curvatures of F and B, respectively. Since these objects are *functions* on the surfaces F and B, they may be unambiguously lifted to M as well.

Proposition 3.79. *Let $M = B \times_f F$ be a 4-dimensional semi-Riemannian warped product with $\dim B = \dim F = 2$ and root warping function $S = \sqrt{f}$. For M to be Ricci flat, it is necessary and sufficient that*

(1) *F have constant sectional curvature ${}^FK$,*
(2) $\frac{\Delta(S^2)}{2} = S\Delta S + g_B(\operatorname{grad} S, \operatorname{grad} S) = {}^FK$ *on B, where ${}^FK$ is the constant value from (1) and Δ denotes the Laplacian on B, and*
(3) $D_X(\operatorname{grad} S) = \left(\frac{{}^BK \cdot S}{2}\right) X$ *for all $X \in \mathcal{L}(B)$.*

Further, if M is Ricci flat, then

(4) $\frac{\Delta S}{S} = {}^BK$ *on B,*

and the sectional curvatures and root warping function are related as follows:

(5) ${}^BK \cdot S^3 = 2C_M$ *on B, where C_M is a constant, and*
(6) ${}^FK = {}^BK \cdot S^2 + g_B(\operatorname{grad} S, \operatorname{grad} S)$ *on B.*

Proof. Using Proposition 3.76, we see that M is Ricci flat if and only if

$$ {}^B\operatorname{Ric}(X,Y) - \tfrac{2}{S}H^S(X,Y) = 0 \quad \text{and} \tag{3.36}$$

$$ {}^F\operatorname{Ric}(V,W) - \langle V,W\rangle S^\sharp = 0 \tag{3.37}$$

for all $X, Y \in \mathfrak{L}(B)$ and $V, W \in \mathfrak{L}(F)$, where $S^\sharp = \frac{\Delta S}{S} + \frac{\langle \operatorname{grad} S, \operatorname{grad} S\rangle}{S^2}$ and Δ denotes the Laplacian on B.

On the semi-Riemannian *surface* B we have

$$ {}^B\operatorname{Ric} = {}^BK \cdot g_B, $$

with the analogous formula holding on the surface F (here the symbol ${}^B\operatorname{Ric}$ denotes the Ricci tensor on B and not its lift to M). Projecting equations (3.36) and (3.37) onto B and F respectively, we see that M is Ricci flat if and only if the following two conditions hold:

$$ {}^{B}K \cdot g_B(X,Y) = \tfrac{2}{S} H^S(X,Y), \qquad \text{and} \tag{3.38} $$

$$ {}^{F}K \cdot g_F(V,W) = S^2 g_F(V,W) S^\sharp \tag{3.39} $$

for all $X, Y \in \mathfrak{X}(B)$ and $V, W \in \mathfrak{X}(F)$. Equation (3.39) must hold as we project from each fiber $p \times F$ into F. Since S is constant on fibers we see that equation (3.39) will hold if and only if

$$ \begin{aligned} {}^{F}K &= \text{ constant, and} \\ f^2 S^\sharp &= S\Delta S + g_B(\operatorname{grad} S, \operatorname{grad} S) = {}^{F}K \qquad \text{on } B. \end{aligned} $$

Rewriting the Hessian as $H^S(X,Y) = g_B(\nabla_X(\operatorname{grad} S), Y)$, equation (3.38) is equivalent to

$$ g_B({}^{B}K \cdot X - \tfrac{2}{S}\nabla_X(\operatorname{grad} S), Y) = 0 \qquad \text{for all } X, Y \in \mathfrak{X}(B) $$

which in turn holds if and only if

$$ \nabla_X(\operatorname{grad} S) = \left({}^{B}K \cdot \tfrac{S}{2}\right) X \qquad \text{for all } X \in \mathfrak{X}(B). $$

That conditions (1), (2) and (3) are necessary and sufficient for M to be Ricci flat now follows. Proposition 3.76–(2) is used to obtain the lifted form of condition (3).

From condition (3) it follows that for a local orthonormal frame field E_0, E_1 on B with $\epsilon_i = g_B(E_i, E_i)$,

$$ \operatorname{div}(\operatorname{grad} S) = \sum_{i=0}^{1} \epsilon_i g_B(\nabla_{E_i}(\operatorname{grad} S), E_i) = \sum_{i=0}^{1} \epsilon_i \left({}^{B}K \cdot \frac{S}{2}\right) g_B(E_i, E_i). $$

Thus the Laplacian of S is given by

$$ \Delta S = {}^{B}K \cdot S \qquad \text{on } B. \tag{3.40} $$

By combining (3.40) with condition (2), we see that

$$ {}^{F}K = {}^{B}K S^2 + g_B(\operatorname{grad} S, \operatorname{grad} S) \tag{3.41} $$

where ${}^F K$ denotes the *constant* value of the sectional curvature of F. Differentiating both sides of equation (3.41) produces

$$\begin{aligned} 0 &= S^2(X\,{}^B K) + {}^B K 2SXS + 2g_B(\nabla_X(\operatorname{grad} S), \operatorname{grad} S) \\ &= S^2(X\,{}^B K) + {}^B K 2SXS + S \cdot {}^B K(XS) \\ &= S^2(X\,{}^B K) + 3 \cdot {}^B K S(XS). \end{aligned}$$

Thus,

$$X({}^B K) = -\left(\frac{3 \cdot {}^B K(XS)}{S}\right) \qquad \text{for all } X \in \mathfrak{X}(B)$$

yielding

$$X(S^3 \cdot {}^B K) = 3S^2(XS)\,{}^B K - S^3\left(\frac{3 \cdot {}^B K(XS)}{S}\right) = 0,$$

and this holds for all lifts $X \in \mathfrak{L}(B)$. It follows that $S^3 \cdot {}^B K$ is constant on B (and hence its lift is constant on M). □

It is now possible to characterize all (2×2) Ricci flat warped products $M = B \times_f F$ with base B of constant curvature.

Corollary 3.80. *If $M = B \times_f F$ is a four-dimensional Ricci flat warped product with base B a surface of constant curvature, then M is simply a product manifold $B \times G$ where B and G are flat two-manifolds. Thus the metric on M is semi-Euclidean.*

Proof. Assume B has constant curvature. Proposition 3.79–(5) shows that the root warping function S, and hence also the warping function f, must be constant on B, and thus M may be viewed as a product manifold. Equations (2) and (4) of Proposition 3.79 now imply that F and B are flat. □

The following result deals with (2×2) warped products $B \times_f F$ where the curvature ${}^B K$ is not constant.

Proposition 3.81. *Let $M = B \times_f F$ be a Ricci flat semi-Riemannian warped product with* $\dim B = \dim F = 2$, *and assume that the sectional curvature ${}^B K$ of the base B is not constant. If* $\operatorname{grad} S$ *is nonnull and* $(\operatorname{grad} S)|_p \neq 0$

at a given point $p \in B$, then there exist local coordinates (t, r) on a neighborhood $\mathcal{U} \subseteq B$ of p such that the following conditions hold on $\mathcal{U}$.

(1) *The metric on B has coordinate expression $ds^2 = E(r)dt^2 + G(r)dr^2$, where $G(r) = \left({}^F K - \frac{2C_M}{r}\right)^{-1}$, $E(r) = \pm G^{-1}$, C_M is a constant, and ${}^F K$ is the constant value of the sectional curvature on F.*

(2) *The root warping function has the form $S(t, r) = r$.*

(3) *The sectional curvature on B is given by ${}^B K(r, t) = \frac{2C_M}{r^3}$.*

The sign of E in condition (1) depends upon the signature of B.

Proof. Assume grad S is nonnull and $(\operatorname{grad} S)|_p \neq 0$. By continuity, we may find a neighborhood $\mathcal{U}$ of p on which the gradient of S is non-vanishing. Let $c_0 = S^{-1}(k) \cap \mathcal{U}$, $k \in \mathbb{R}^+$, denote a single level curve of S in $\mathcal{U}$. Consider geodesics intersecting c_0 orthogonally; these geodesics are also integral curves of grad S, and the orthogonal trajectories of these geodesics are level sets of S. Applying the classical geodesic coordinate construction, we introduce local coordinates (v, u) such that the geodesics are the $v =$ constant curves and the orthogonal trajectories are the $u =$ constant curves. In these coordinates the metric assumes the form

$$ds^2 = H(u, v)dv^2 + F(u)du^2.$$

Rescale coordinates, introducing $r = r(u)$ and $t = t(v)$ such that

$$r(u) = S(u) \quad \text{and} \quad D_{\partial/\partial t}(\partial/\partial t) = 0.$$

Thus, r merely traces the values of the warping function S, and t is an affine coordinate along the geodesics. It should be noted that this will imply $r > 0$. In (t, r) coordinates we have metric tensor components

$$ds^2 = E(r, t)dt^2 + G(r)dr^2$$

with the root warping function given simply by

$$S(t, r) = r.$$

Furthermore, since

$$\operatorname{grad} S = \sum_{i,j} g^{ij} \frac{\partial S}{\partial x^i} \frac{\partial}{\partial x^j},$$

the gradient of S is given in (r,t) coordinates by

$$\operatorname{grad} S = \frac{1}{G(r)} \frac{\partial}{\partial r}.$$

Equation (5) of Proposition 3.79 now shows that the curvature on the surface B is a function solely of r and is given by

$${}^{B}K(r,t) = \frac{2C_M}{[S(t,r)]^3} = \frac{2C_M}{r^3}.$$

Now $g_B(\operatorname{grad} S, \operatorname{grad} S) = \frac{1}{G}$, so formula (6) of Proposition 3.79 yields

$$\begin{aligned} {}^{F}K &= {}^{B}K \cdot S^2 + \frac{1}{G(r)} \\ &= \frac{2C_M}{S^3} S^2 + \frac{1}{G(r)} \\ &= \frac{2C_M}{r} + \frac{1}{G(r)}. \end{aligned}$$

We have determined the metric coefficient G:

$$G(r) = \left({}^{F}K - \frac{2C_M}{r} \right)^{-1}$$

It remains to determine the form of the metric coefficient $E(t,r)$. We first show that E is a function only of the variable r and then derive the function $E(r)$.

Note that

$$D_{\partial/\partial t}(\partial/\partial t) = 0 = \Gamma_{11}^{1}(\partial/\partial t) + \Gamma_{11}^{2}(\partial/\partial r),$$

so that in particular $\Gamma_{11}^{1} = 0$. However, since (t,r) are orthogonal coordinates, we have

$$E \cdot G \cdot \Gamma_{11}^{1} = \frac{1}{2} E_t \cdot G.$$

It follows that $E_t = 0$ and E is a function solely of r.

Now using Proposition 3.79 with $\operatorname{grad} S = \frac{1}{G(r)}\frac{\partial}{\partial r}$, we obtain

$$D_{\partial/\partial t}\left(\frac{1}{G(r)}\frac{\partial}{\partial r}\right) = \left({}^{B}K \cdot \frac{S}{2}\right)\frac{\partial}{\partial t}, \quad \text{or equivalently}$$

$$\frac{1}{G}D_{\partial/\partial t}(\partial/\partial r) = \frac{C_M}{r^2}\frac{\partial}{\partial t}, \quad \text{giving}$$

$$D_{\partial/\partial t}(\partial/\partial r) = \frac{G \cdot C_M}{r^2}\frac{\partial}{\partial t}.$$

However, in the orthogonal (t, r) coordinate system, we have

$$D_{\partial/\partial t}(\partial/\partial r) = \frac{E_r}{2E}\frac{\partial}{\partial t} + \frac{G_t}{2G}\frac{\partial}{\partial r} = \frac{E_r}{2E}\frac{\partial}{\partial t}.$$

It follows that

$$\frac{E_r}{E} = G(r)\frac{2C_M}{r^2}, \quad \text{and thus}$$

$$\frac{d(\ln|E|)}{dr} = \pm\left({}^{F}K - \frac{2C_M}{r}\right)^{-1}\left(\frac{2C_M}{r^2}\right), \quad \text{giving}$$

$$\ln|E| = \pm\int \frac{\left(\frac{2C_M}{r^2}\right)}{\left({}^{F}K - \frac{2C_M}{r}\right)}dr.$$

With the substitution $y = {}^{F}K - \frac{2C_M}{r}$, this becomes

$$\ln|E| = \pm\int\frac{dy}{y}, \quad \text{giving}$$

$$\ln|E| = \pm\ln|y| + c, \quad \text{or finally}$$

$$E = \pm a \cdot \left({}^{F}K - \frac{2C_M}{r}\right)$$

where $a = e^c$ is positive. In orthogonal coordinates with $e = \sqrt{|E|}$ and $g = \sqrt{|G|}$, the curvature is given by the classical formula

$${}^{B}K = \frac{-1}{eg}\left[\epsilon_t\left(\frac{g_t}{e}\right)_t + \epsilon_r\left(\frac{e_r}{g}\right)_r\right].$$

Since E and G are functions of r only, this reduces to the following simple formula. We are ignoring the case-by-case analysis of the signs involving the absolute values since the end result will show such considerations to be unnecessary.

$$\begin{aligned} {}^BK &= \pm\frac{1}{\sqrt{a}}\left[\epsilon_r \frac{\frac{1}{2}a({}^FK - \frac{2C_M}{r})^{-\frac{1}{2}}\left(\frac{2C_M}{r^2}\right)}{({}^FK - \frac{2C_M}{r})^{-\frac{1}{2}}}\right]_r \\ &= \pm\epsilon_r\sqrt{a}\left[\frac{C_M}{r^2}\right]_r \\ &= \pm\frac{2\epsilon_r\sqrt{a}C_M}{r^3}. \end{aligned}$$

Since ${}^BK = \frac{2C_M}{r^3}$, we see that $a = 1$, and the sign in this final term is determined by the signature of B and the sign of $G(r)$. We have therefore shown that

$$E(r) = \pm\left({}^FK - \frac{2C_M}{r}\right) = \pm G(r)^{-1}. \quad \square$$

Observe that ∂_t is a local nonvanishing Killing field on $\mathcal{U}$ in the above lemma, which is valid for spaces of arbitrary signature. Let us now fix the signature of all spaces under consideration to the Lorentzian signature $(-,+,+,+)$. Corollary 3.80 and Proposition 3.81 now have the following import: a Ricci flat (2×2) warped product M is completely specified (locally) by the choice of base (B, g_B) and the constants ${}^FK$ and C_M in Proposition 3.79. Since we are working *locally*, we may assume the fiber F of constant sectional curvature ${}^FK$ is a subset of

(1) the sphere $S^2(\rho)$ if ${}^FK = \frac{1}{\rho^2}$,
(2) Euclidean space $\mathbb{R}^2$ if ${}^FK = 0$, or
(3) hyperbolic space $H^2(\rho)$ if ${}^FK = -\frac{1}{\rho^2}$.

Theorem 3.82. *Assume (M, g) is a Ricci flat four-manifold of Lorentzian signature $(-,+,+,+)$, and let $p \in M$ be any point of M. The point p has a neighborhood $\mathcal{U}$ such that $(\mathcal{U}, g|_{\mathcal{U}})$ is isometric to a (2×2) warped product with both parameters $(C_M, {}^FK)$ positive and nonconstant sectional curvature ${}^BK$ on B if and only if M is locally isometric to an open subset of Schwarzschild space–time with mass $M_0 = \dfrac{C_M}{({}^FK)^{\frac{3}{2}}}$.*

Proof. Assume M admits such a local warped product structure. Since we are working locally, we may assume the fiber F is the 2-sphere of radius $\frac{1}{\sqrt{{}^FK}}$.

Let $d\sigma^2$ denote the canonical metric on the unit 2-sphere. Proposition 3.81 implies the metric may be expressed on $\mathcal{U}$ in the form

$$ds^2 = -E(r)dt^2 + E(r)^{-1}dr^2 + r^2 \frac{1}{{}^FK} d\sigma^2$$

with $E(r) = \left({}^FK - \frac{2C_M}{r}\right)$.

Upon making the change of variables $u = \frac{r}{\sqrt{{}^FK}}$ and $v = t\sqrt{{}^FK}$, the metric becomes

$$ds^2 = -\left(1 - \frac{\left[2C_M/({}^FK)^{\frac{3}{2}}\right]}{u}\right) dv^2 + \left(1 - \frac{\left[2C_M/({}^FK)^{\frac{3}{2}}\right]}{u}\right) du^2 + u^2 d\sigma^2,$$

thus demonstrating the first claim. The converse is now clear. □

This result should be contrasted with Birkhoff's Theorem [cf. Hawking and Ellis (1973, p. 372)] on spherically symmetric solutions to Einstein's vacuum field equations, which also offers an alternate proof of Theorem 3.82.

Theorem (Birkhoff). *Any C^2 solution of Einstein's empty space equations which is spherically symmetric in an open set V is locally equivalent to part of the maximally extended Schwarzschild solution in V.*

The typical construction of Schwarzschild space–time [cf. O'Neill (1983)] follows the proof of Birkhoff's theorem. A number of strong physical assumptions, not the least of which is spherical symmetry, lead one inevitably to Schwarzschild space–time as the unique model. What we have shown is that a Ricci flat space–time which possesses enough symmetry to be expressed locally as a (2×2) warped product must have spherical, planar, or hyperbolic symmetry. In the first case we obtain the conclusion to Birkhoff's Theorem: M is locally isometric to a portion of Schwarzschild space–time.

The following result may be established by an argument similar to the one employed in the proof of Proposition 3.79 [cf. Easley (1991)].

Theorem 3.83. *Let $M = B \times_f F$ be a semi-Riemannian warped product with $\dim B = 1$ and $\dim F = n \geq 2$. For M to be Ricci flat, it is necessary and sufficient that the following two conditions hold.*

(1) *The root warping function S satisfies $\langle \operatorname{grad} S, \operatorname{grad} S \rangle = C_0$, a constant.*

(2) *(F, g_F) is an Einstein manifold with* ${}^F\mathrm{Ric} = 2\langle \operatorname{grad} S, \operatorname{grad} S\rangle g_F = C_0 \cdot g_F$.

This result indicates that a Ricci flat manifold (M, g) can admit only one of a restricted class of local warped product splittings at $p \in M$ if we require $\dim B = 1$. If $\dim F = 2$, for example, it follows that F is a surface of *constant curvature* ${}^F K$ since ${}^F\mathrm{Ric} = {}^F K g_F$ in this case. If we also assume ${}^F K = 0$, we have $S = \sqrt{f} =$ constant, and the warped product is merely a standard product manifold with a semi-Euclidean metric.

If $\dim F = 3$, it follows that the Einstein manifold (F, g_F) has *constant curvature* [cf. Petrov (1969, p. 77)]. If (F, g_F) is Riemannian, we may consider F a subset of

(1) the sphere $S^3(r)$ if ${}^F K = \frac{1}{r^2}$,
(2) Euclidean space $\mathbb{R}^3$ if ${}^F K = 0$, or
(3) hyperbolic space $H^3(r)$ if ${}^F K = -\frac{1}{r^2}$,

with analogous restrictions holding if g_F is indefinite. An exact solution to Einstein's equations which uses $F = S^3$ is the spatially homogeneous Taub-NUT model [cf. Hawking and Ellis (1973, p. 170)].

CHAPTER 4

LORENTZIAN DISTANCE

With the basic properties of Riemannian metrics in mind (cf. Chapter 1), it is the aim of this chapter to study the corresponding properties of Lorentzian distance and to show how the Lorentzian distance is related to the causal structure of the given space–time. We also show that Lorentzian distance preserving maps of a strongly causal space–time onto itself are diffeomorphisms which preserve the metric tensor.

While there are many similarities between the Riemannian and Lorentzian distance functions, many basic differences will also be apparent from this chapter. Nonetheless, a duality between "minimal" for Riemannian manifolds and "maximal" for Lorentzian manifolds will be noticed in this and subsequent chapters.

4.1 Basic Concepts and Definitions

Let (M, g) be a Lorentzian manifold of dimension $n \geq 2$. Given $p, q \in M$ with $p \leqslant q$, let $\Omega_{p,q}$ denote the path space of all future directed nonspacelike curves $\gamma : [0, 1] \to M$ with $\gamma(0) = p$ and $\gamma(1) = q$. The Lorentzian arc length functional $L = L_g : \Omega_{p,q} \to \mathbb{R}$ is then defined as follows [cf. Hawking and Ellis (1973, p. 105)]. Given a piecewise smooth curve $\gamma \in \Omega_{p,q}$, choose a partition $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1$ such that $\gamma \mid (t_i, t_{i+1})$ is smooth for each $i = 0, 1, \ldots, n-1$. Then define

$$(4.1) \qquad L(\gamma) = L_g(\gamma) = \sum_{i=0}^{n-1} \int_{t=t_i}^{t_{i+1}} \sqrt{-g(\gamma'(t), \gamma'(t))}\, dt.$$

It may be checked as in elementary differential geometry [cf. O'Neill (1966, pp. 51–52)] that this definition of Lorentzian arc length is independent of

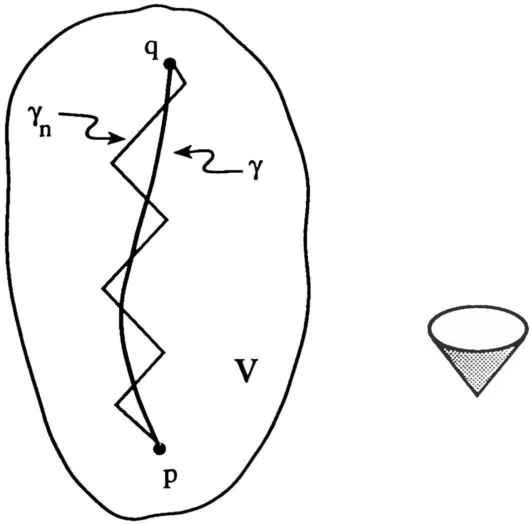


FIGURE 4.1. The timelike curve γ from p to q is approximated by a sequence of curves $\{\gamma_n\}$ with $\gamma_n \to \gamma$ in the C^0 topology but $L(\gamma_n) \to 0$.

the parametrization of γ. Since an arbitrary nonspacelike curve satisfies a local Lipschitz condition, it is differentiable almost everywhere. Hence the Lorentzian arc length $L(\gamma)$ of γ may still be defined using (4.1). Alternative but equivalent definitions of $L(\gamma)$ for arbitrary nonnull nonspacelike curves may be given by approximating γ by C^1 timelike curves [cf. Hawking and Ellis (1973, p. 214)] or by approximating γ by sequences of broken nonspacelike geodesics [cf. Penrose (1972, p. 53)]. The Lorentzian arc length of an arbitrary null curve may be set equal to zero.

Now fix $p, q \in M$ with $p \ll q$. If γ is any timelike curve from p to q, then $L(\gamma) > 0$. On the other hand, γ may be approximated by a sequence $\{\gamma_n\}$ of piecewise smooth "almost null" curves $\gamma_n : [0, 1] \to M$ with $\gamma_n(0) = p$ and $\gamma_n(1) = q$ such that $\gamma_n \to \gamma$ in the C^0 topology, but $L(\gamma_n) \to 0$ (cf. Figure 4.1). This construction shows, moreover, that given any $p, q \in M$ with $p \ll q$, there are curves $\gamma \in \Omega_{p,q}$ with arbitrarily small Lorentzian arc length. Hence, the infimum of Lorentzian arc lengths of all piecewise smooth curves joining any two chronologically related points $p \ll q$ is always zero. However, if p and q lie in a geodesically convex neighborhood U, the future directed timelike geodesic segment in U from p to q has the largest Lorentzian arc length among all nonspacelike curves in U from p to q. Thus it is natural to make the following definition of the Lorentzian distance function $d = d(g) : M \times M \to \mathbb{R} \cup \{\infty\}$ of (M, g).

Definition 4.1. (*Lorentzian Distance Function*) Given a point p in M, if $q \notin J^+(p)$, set $d(p, q) = 0$. If $q \in J^+(p)$, set $d(p, q) = \sup\{L_g(\gamma) : \gamma \in \Omega_{p,q}\}$.

From the definition, it is immediate that

$$d(p, q) > 0 \quad \text{if and only if} \quad q \in I^+(p). \tag{4.2}$$

Thus the Lorentzian distance function determines the chronological past and future of any point. However, the Lorentzian distance function in general fails to determine the causal past and future sets of p since $d(p, q) = 0$ does *not* imply $q \in J^+(p) - I^+(p)$. But at least if $q \in J^+(p) - I^+(p)$, then $d(p, q) = 0$.

We emphasize that the Lorentzian distance $d(p, q)$ need *not* be finite. One way that the condition $d(p, q) = \infty$ may occur is that timelike curves from p to q may attain arbitrarily large arc lengths by approaching certain boundary points of the space–time. In Figure 4.2, two points with $d(p, q) = \infty$ are shown in a Reissner–Nordström space–time with $e^2 = m^2$ [cf. Hawking and Ellis (1973, p. 160)].

A second way Lorentzian distance may become infinite is through causality violations. Recall that a space–time is said to be vicious at the point $p \in M$ if $I^+(p) \cap I^-(p) = M$ and totally vicious if $I^+(p) \cap I^-(p) = M$ for all $p \in M$.

Lemma 4.2. *Let (M, g) be an arbitrary space–time.*

(1) *If $p \in I^+(p)$, then $d(p, p) = \infty$. Thus for each $p \in M$, either $d(p, p) = 0$ or $d(p, p) = \infty$.*
(2) *(M, g) is totally vicious iff $d(p, q) = \infty$ for all $p, q \in M$.*
(3) *If (M, g) is vicious at p, then (M, g) is totally vicious.*

Proof. (1) Suppose $p \in I^+(p)$. Then we may find a closed timelike curve $\gamma : [0,1] \to M$ with $\gamma(0) = \gamma(1) = p$. Since γ is timelike, $L(\gamma) > 0$. If $\sigma_n \in \Omega_{p,p}$ is the timelike curve obtained by traversing γ exactly n times, then $L(\sigma_n) = nL(\gamma) \to \infty$ as $n \to \infty$. Thus $d(p,p) = \infty$.
(2) Suppose (M, g) is totally vicious. Fix $p, q \in M$, and let $n > 0$ be any positive integer. Since $p \in I^+(p)$, we may find $\gamma_1 \in \Omega_{p,p}$ with $L(\gamma_1) \geq n$ by part (1). Since $q \in I^+(p)$, there is a timelike curve γ_2 from p to q. Then $\gamma = \gamma_1 * \gamma_2 \in \Omega_{p,q}$ is a timelike curve with length $L(\gamma) = L(\gamma_1) + L(\gamma_2) > n$. Hence $d(p,q) = \infty$.

Conversely, suppose $d(p,q) = \infty$ for all $p, q \in M$. Fixing $r \in M$, we have $d(r,p) > 0$ and $d(p,r) > 0$ for all $p \in M$. Thus by (4.2), it follows that $I^+(r) \cap I^-(r) = M$.
(3) Inspired by Lemma 4.2–(2), T. Ikawa and H. Nakagawa (1988) proved (3) using the Lorentzian distance function. Subsequently, B. Wegner (1989) noted that the following elementary argument yields the desired result. Let $q \in M = I^+(p) \cap I^-(p)$. Then $q \in I^+(p)$ so $I^-(p) \subseteq I^-(q)$ by the transitivity of $\ll$. Similarly, $q \in I^-(p)$ implies $I^+(p) \subseteq I^+(q)$. Hence, $M = I^+(p) \cap I^-(p) \subseteq I^+(q) \cap I^-(q)$. □

By Definition 4.1, if $I^+(p) \neq M$ then there are points $q \in M$ with $d(p,q) = 0$ but $p \neq q$. Hence unlike the Riemannian distance function, the Lorentzian distance function usually fails to be nondegenerate. Indeed, we have seen that $d(p,p) > 0$ is possible. But if (M, g) is chronological, then $d(p,p) = 0$ for all $p \in M$. Also, the Lorentzian distance function tends to be nonsymmetric. More precisely, the following may be shown for arbitrary space–times.

Remark 4.3. If $p \neq q$ and $d(p,q)$ and $d(q,p)$ are both finite, then either $d(p,q) = 0$ or $d(q,p) = 0$. Equivalently, if $d(p,q) > 0$ and $d(q,p) > 0$, then $d(p,q) = d(q,p) = \infty$.

Proof. If $d(p,q) > 0$ and $d(q,p) > 0$, we may find future directed timelike curves γ_1 from p to q and γ_2 from q to p, respectively. Define a sequence $\{\gamma_n\}$ by $\gamma_n = \gamma_1 * (\gamma_2 * \gamma_1)^n \in \Omega_{p,q}$. As $n \to \infty$, $L(\gamma_n) \to \infty$, whence $d(p,q)$ is infinite. Similarly, $d(q,p) = \infty$. □

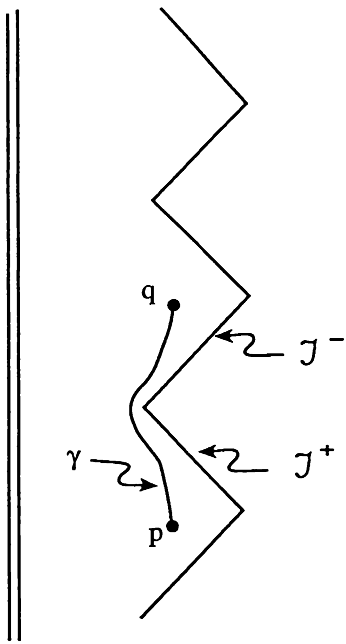


FIGURE 4.2. A Reissner–Nordström space–time with $e^2 = m^2$ is shown. By taking timelike curves γ from p to q close to $\mathfrak{I}^+$ and $\mathfrak{I}^-$, we can make $L(\gamma)$ arbitrarily large. Thus $d(p,q) = \infty$, which means an accelerated observer may take arbitrarily large amounts of time in going from p to q.

A further consequence of Definition 4.1 is that if $\gamma : [0,\infty) \to (M,g)$ is any future directed, future complete, timelike geodesic in an arbitrary space–time (M,g), then $\lim_{t\to\infty} d(\gamma(0),\gamma(t)) \geq \lim_{t\to\infty} L(\gamma\,|\,[0,t]) = \infty$. By contrast, complete Riemannian manifolds (N,g_0) may contain (nonclosed) geodesics $\sigma : [0,\infty) \to (N,g_0)$ for which $\sup\{d_0(\sigma(0),\sigma(t)) : t \geq 0\}$ is finite. Further assumptions are needed for Riemannian manifolds to guarantee that $\lim_{t\to\infty} d(\sigma(0),\sigma(t)) = \infty$ for all geodesics $\sigma : [0,\infty) \to (N,g_0)$ [cf. Cheeger and Ebin (1975, pp. 53 and 151)].

While the Lorentzian distance function fails to be symmetric and nondegenerate, at least a reverse triangle inequality holds (cf. Figure 1.3). Explicitly,

(4.3) If $p \leqslant r \leqslant q$, then $d(p,q) \geq d(p,r) + d(r,q)$.

We now discuss some properties of the Lorentzian distance that make it a useful tool in general relativity and Lorentzian geometry. First, the Lorentzian distance function is lower semicontinuous where it is finite [cf. Hawking and Ellis (1973, p. 215)].

Lemma 4.4. *For Lorentzian distance d, if $d(p,q) < \infty$, $p_n \to p$, and $q_n \to q$, then $d(p,q) \leq \liminf d(p_n, q_n)$. Also, if $d(p,q) = \infty$, $p_n \to p$, and $q_n \to q$, then $\lim_{n\to\infty} d(p_n, q_n) = \infty$.*

Proof. First consider the case $d(p,q) < \infty$. If $d(p,q) = 0$, there is nothing to prove. If $d(p,q) > 0$, then $q \in I^+(p)$ and the lower semicontinuity follows from the following fact. Given any $\epsilon > 0$, a timelike curve γ of length $L \geq d(p,q) - \epsilon/2$ from p to q and sufficiently small neighborhoods U_1 of p and U_2 of q may be found such that γ may be deformed to give a timelike curve of length $L' \geq d(p,q) - \epsilon$ from any point r of U_1 to any point s of U_2.

Suppose now that $d(p,q) = \infty$ but $\liminf d(p_n, q_n) = R < \infty$. Since $d(p,q) = \infty$ there exists a timelike curve γ from p to q of length $L(\gamma) > R + 2$. This implies that there exist neighborhoods U_1 and U_2 of p and q, respectively, such that γ can be deformed to give a timelike curve of length $L' \geq R + 1$ from any point r of U_1 to any point s of U_2. This contradicts $\liminf d(p_n, q_n) = R$. □

In general, the Lorentzian distance function fails to be upper semicontinuous. We give an example of a space–time (M, g) containing an infinite sequence $\{p_n\}$ with $p_n \to p$ and a point $q \in I^+(p)$ such that $d(p_n, q) = \infty$ for all large n but $d(p,q) < \infty$ (cf. Figure 4.3).

For globally hyperbolic space–times, on the other hand, the Lorentzian distance function is finite and continuous just like the Riemannian distance function.

Lemma 4.5. *For a globally hyperbolic space–time (M, g), the Lorentzian distance function d is finite and continuous on $M \times M$.*

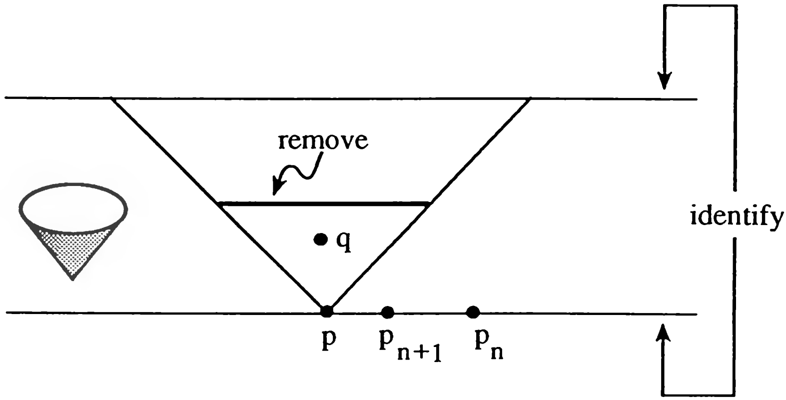


FIGURE 4.3. Let M be $\{(x,y) : 0 \le y \le 2\} - \{(x,1) : -1 \le x \le 1\}$ with the identification $(x,0) \sim (x,2)$ and the flat Lorentzian metric $ds^2 = dx^2 - dy^2$. Let $p = (0,0)$, $q = (0,1/2)$, and $p_n \to p$ as shown. Then $p_n \in I^+(p_n)$ and hence $d(p_n, p_n) = \infty$ for all n. For large n we have $q \in I^+(p_n)$ and thus $d(p_n, q) = \infty$. On the other hand, $d(p,q) = 1/2$ which yields $d(p,q) < \liminf d(p_n, q)$. This space–time is not causal. However, the distance function may also fail to be upper semicontinuous in causal space–times (cf. Figure 4.6).

Proof. To prove the finiteness of d, cover the compact set $J^+(p) \cap J^-(q)$ with a finite number of convex normal neighborhoods $B_1, B_2, \ldots, B_m$ such that no nonspacelike curve which leaves any B_i ever returns and such that every nonspacelike curve in each B_i has length at most one. Since any nonspacelike curve γ from p to q can enter each B_i no more than once, $L(\gamma) \le m$. Hence $d(p,q) \le m$.

If d failed to be upper semicontinuous at $(p,q) \in M \times M$, we could find a $\delta > 0$ and sequences $\{p_n\}$ and $\{q_n\}$ converging to p and q respectively, such that $d(p_n, q_n) \ge d(p,q) + 2\delta$ for all n. By definition of $d(p_n, q_n)$, we may then find a future directed nonspacelike curve γ_n from p_n to q_n with $L(\gamma_n) \ge d(p,q) + \delta$ for each n. By Corollary 3.32, the sequence $\{\gamma_n\}$ has a nonspacelike limit curve γ from p to q. By Proposition 3.34, a subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$ converges to γ in the C^0 topology. Hence $L(\gamma) \geq d(p,q) + \delta$ by Remark 3.35. But this contradicts the definition of Lorentzian distance. Thus d is upper semicontinuous at (p,q). □

We now define the following distance condition [cf. Beem and Ehrlich (1977, Condition 4)].

Definition 4.6. (*Finite Distance Condition*) The space–time (M,g) satisfies the *finite distance condition* if $d(g)(p,q) < \infty$ for all $p, q \in M$.

Lemma 4.5 then has the following corollary.

Corollary 4.7. *If (M,g) is globally hyperbolic, then (M,g) satisfies the finite distance condition and $d(g) : M \times M \to \mathbb{R}$ is continuous.*

If (M,g) is globally hyperbolic, all metrics in the conformal class $C(M,g)$ are globally hyperbolic. Hence all metrics in $C(M,g)$ satisfy the finite distance condition. We will examine the converse of this statement in Section 4.3, Theorem 4.30.

Since the given topology of a smooth manifold coincides with the metric topology induced by any Riemannian metric, it is natural to consider the sets $\{m \in I^+(p) : d(p,m) < \epsilon\}$ for a Lorentzian manifold. However, as Minkowski space shows, these sets do not form a basis for the given manifold topology (cf. Figure 4.4). Indeed, this same example shows that no matter how small $\epsilon > 0$ is chosen, the sets $\{m \in J^+(p) : d(p,m) \leq \epsilon\}$ may fail to be compact and fail to be geodesically convex as well as failing to be diffeomorphic to the closed n-disk.

The sphere of radius ϵ for the point $p \in M$ is given by $K(p,\epsilon) = \{q \in M : d(p,q) = \epsilon\}$. This set need not be compact. However, the reverse triangle inequality and (4.2) imply that $K(p,\epsilon)$ is achronal for all finite $\epsilon > 0$ and all $p \in M$.

In arbitrary space–times, neither the future inner ball

$$B^+(p,\epsilon) = \{q \in I^+(p) : d(p,q) < \epsilon\}$$

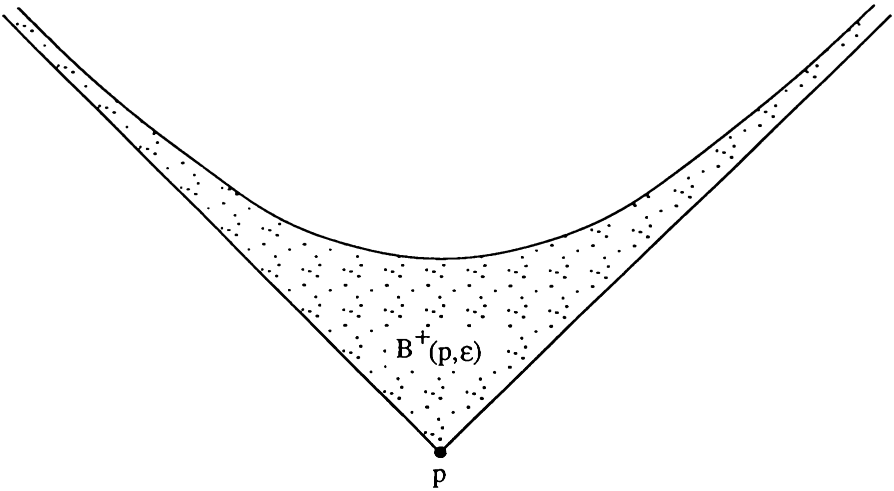


FIGURE 4.4. The set $B^+(p, \epsilon) = \{q \in I^+(p) : d(p, q) < \epsilon\}$ in Minkowski space–time does not have compact closure, is not geodesically convex, and does not contain p. Furthermore, sets of the form $B^+(p, \epsilon)$ do not form a basis for the manifold topology. But in general, if (M, g) is a distinguishing space–time with a continuous Lorentzian distance function, then a subbasis for the manifold topology is given by sets of the form $B^+(p, \epsilon)$ and $B^-(p, \epsilon)$ [cf. Proposition 4.31]. Hence these sets do form a *subbasis* for the given topology of Minkowski space–time.

nor the past inner ball

$$B^-(p, \epsilon) = \{q \in I^-(p) : d(q, p) < \epsilon\}$$

need be open. On the other hand, when the distance function $d : M \times M \rightarrow \mathbb{R} \cup \{\infty\}$ is continuous, these inner balls must be open. In Section 4.3 we will show that for distinguishing space–times with continuous distance functions, the past and future inner balls form a subbasis for the manifold topology.

A different subbasis for the topology of any strongly causal space–time (M, g) with a possibly discontinuous distance function $d = d(g) : M \times M \to \mathbb{R} \cup \{\infty\}$ may be obtained by using the outer balls $O^+(p, \epsilon)$ and $O^-(p, \epsilon)$ rather than the inner balls.

Definition 4.8. (*Outer Balls* $O^+(p, \epsilon)$, $O^-(p, \epsilon)$) The *outer ball* $O^+(p, \epsilon)$ [respectively, $O^-(p, \epsilon)$] of $I^+(p)$ [respectively, $I^-(p)$] is given by

$$O^+(p, \epsilon) = \{q \in M : d(p, q) > \epsilon\},$$

respectively,

$$O^-(p, \epsilon) = \{q \in M : d(q, p) > \epsilon\}$$

(cf. Figure 4.5).

Since the Lorentzian distance function is lower semicontinuous where it is finite, the outer balls $O^+(p, \epsilon)$ and $O^-(p, \epsilon)$ are open in arbitrary space–times. The reverse triangle inequality implies that these sets also have the property that if $m, n \in O^+(p, \epsilon)$ [respectively, $m, n \in O^-(p, \epsilon)$] and $m \leqslant n$, then any future directed nonspacelike curve from m to n lies in $O^+(p, \epsilon)$ [respectively, $O^-(p, \epsilon)$]. Moreover, we have

Theorem 4.9. *Let* (M, g) *be strongly causal. Then the collection*

$$\{O^+(p, \epsilon_1) \cap O^-(q, \epsilon_2) : p, q \in M,\ \epsilon_1, \epsilon_2 > 0\}$$

forms a basis for the given manifold topology.

Proof. Let $m \in M$ be given, and let U be any open neighborhood containing m. We may find a local causality neighborhood U_1 with $m \in U_1 \subseteq U$, i.e., no nonspacelike curve which leaves U_1 ever returns. Choose $p_1, p_2 \in U_1$ with $p_1 \ll m \ll p_2$ such that $I^+(p_1) \cap I^-(p_2) \subseteq U_1$. By the chronology assumptions on p_1 and p_2, we have $d(p_1, m) > 0$ and $d(m, p_2) > 0$. Choose constants ϵ_1, ϵ_2 with $0 < \epsilon_1 < d(p_1, m)$ and $0 < \epsilon_2 < d(m, p_2)$. Then $m \in O^+(p_1, \epsilon_1) \cap O^-(p_2, \epsilon_2)$. Since $O^+(p_1, \epsilon_1) \subseteq I^+(p_1)$ and $O^-(p_2, \epsilon_2) \subseteq I^-(p_2)$, we also have $O^+(p_1, \epsilon_1) \cap O^-(p_2, \epsilon_2) \subseteq I^+(p_1) \cap I^-(p_2) \subseteq U_1 \subseteq U$ as required. $\square$

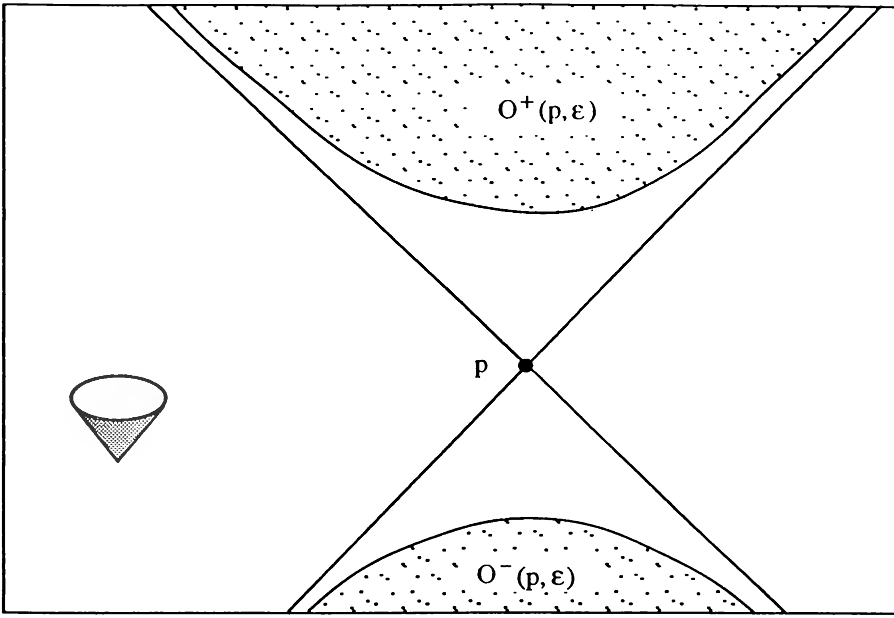


FIGURE 4.5. The outer balls $O^+(p, \epsilon) = \{q \in M : d(p, q) > \epsilon\}$ and $O^-(p, \epsilon) = \{q \in M : d(q, p) > \epsilon\}$ are open in arbitrary space–times. Furthermore, $O^+(p, \epsilon)$ and $O^-(p, \epsilon)$ are always subsets of $I^+(p)$ and $I^-(p)$, respectively. If (M, g) is strongly causal, the outer balls $O^+(p, \epsilon)$ and $O^-(p, \epsilon)$ with $p \in M$ and $\epsilon > 0$ arbitrary form a subbasis for the manifold topology.

For complete Riemannian manifolds, any two points may be joined by a minimal (distance-realizing) geodesic segment. We now examine the dual of this property for space–times.

In Hawking and Ellis (1973, p. 110), a timelike geodesic γ from p to q is said to be maximal if the index form of γ is negative semidefinite. This definition implies that if the geodesic γ is *not* maximal, there exist variations of γ which yield curves from p to q "close" to γ having longer Lorentzian arc length than γ. If γ is maximal in this sense, however, no small variation of γ keeping p and q fixed will produce timelike curves σ from p to q with $L(\sigma) > L(\gamma)$.

Nonetheless, there may still exist a timelike geodesic σ_1 in M from p to q ("far" from γ) with $d(p,q) = L(\sigma_1) > L(\gamma)$. Thus maximality as defined by Hawking and Ellis does not imply "maximality in the large." To study "maximality in the large," we adopt, in analogy to the concept of minimality in Riemannian geometry, a definition of maximality valid for all *curves* in the path space $\Omega_{p,q}$ [cf. Beem and Ehrlich (1977, Definition 1)]. The motivation for our definition is Theorem 4.13 below [cf. Beem and Ehrlich (1979a, p. 166)] and its applications, particularly the construction of geodesics as limit curves of sequences of "almost maximal" curves in Chapter 8 and the definition of the Lorentzian cut locus in Chapter 9.

Definition 4.10. (*Maximal Curve*) Let $p, q \in M$ with $p \leqslant q$, $p \neq q$. The curve $\gamma \in \Omega_{p,q}$ is said to be *maximal* if $L(\gamma) = d(p,q)$.

An immediate consequence of the reverse triangle inequality (4.3) is

Remark 4.11. If $\gamma : [0,1] \to M$ in $\Omega_{p,q}$ is maximal, then for all s, t with $0 \leq s < t \leq 1$, we have $d(\gamma(s), \gamma(t)) = L(\gamma \,|\, [s,t\,])$.

The following result, stated somewhat differently in Penrose (1972, Proposition 7.2), is the analogue of the principle in Riemannian geometry that "locally" geodesics minimize arc length [cf. Bishop and Crittenden (1964, p. 149, Theorem 2)].

Proposition 4.12. *Let U be a convex normal neighborhood centered at point $p \in M$. For $q \in J^+(p)$, let $\overline{pq}$ denote the unique nonspacelike geodesic $c : [0,1] \to U$ in U with $c(0) = p$ and $c(1) = q$. If γ is any future directed nonspacelike curve in U from p to q with $L(\gamma) = d(p,q)$, then γ coincides with $\overline{pq}$ up to parametrization.*

Proof. For $q \in I^+(p)$ and $d(p,q) > 0$, Penrose (1972, p. 53) shows, using a synchronous coordinate system, that if γ is any causal trip in U from p to q other than $\overline{pq}$, then $L(\gamma) < L(\overline{pq}) = d(p,q)$. This may be obtained equivalently using the Gauss Lemma (cf. Corollary 10.19 of Section 10.1). Hence the result is established if $d(p,q) > 0$.

Suppose now that $d(p,q) = 0$, and let γ be any nonspacelike curve in U from p to q. Then $L(\gamma) \leq d(p,q) = 0$. Thus $\gamma : [0,1] \to M$ is a null curve. Suppose that $\dot{\gamma}(t) \notin \mathrm{Int}(\overline{pq})$. Let γ_1 be the unique null geodesic in U from p to $\gamma(t)$, and let γ_2 be the unique null geodesic in U from $\gamma(t)$ to q. By Proposition 2.19 of Penrose (1972, p. 15), $\gamma_1 * \gamma_2$ is either a smooth null geodesic or $p \ll q$. Since $d(p,q) = 0$, $p \ll q$ is impossible. Hence $\gamma_1 * \gamma_2$ is a smooth null geodesic which, by convexity of U, must coincide with $\overline{pq}$ up to parametrization. □

Proposition 4.12 has the following important consequence.

Theorem 4.13. *If $\gamma \in \Omega_{p,q}$ satisfies $L(\gamma) = d(p,q)$, then γ may be reparametrized to be a smooth geodesic.*

Proof. Fix any point $\gamma(t)$ on γ. We may find $\delta > 0$ such that a convex neighborhood centered at $\gamma(t+\delta)$ contains $\gamma([t-\delta, t+\delta])$. By Remark 4.11 the curve $\gamma \mid [t-\delta, t+\delta]$ is maximal. Thus Proposition 4.12 implies that $\gamma \mid [t-\delta, t+\delta]$ may be reparametrized to be a smooth geodesic. As t was arbitrary, the theorem now follows. □

As an illustration of the use of Definition 4.10 and Theorem 4.13, we give a simple proof of a basic result in elementary causality theory [cf. Penrose (1972, Proposition 2.20)] that is usually obtained by different methods.

Corollary 4.14. *If $p \leqslant q$ but it is not the case that $p \ll q$, then there is a maximal null geodesic from p to q.*

Proof. The causality assumptions on p and q imply that $d(p,q) = 0$. Now let γ be a future directed nonspacelike curve from p to q. By definition of Lorentzian distance, $d(p,q) \geq L(\gamma) \geq 0$. Thus $L(\gamma) = d(p,q) = 0$ and γ is maximal. By Theorem 4.13, γ may be reparametrized to a smooth geodesic $c : [0,1] \to M$ from p to q. Since $L(c) \leq d(p,q) = 0$, the geodesic c must be a null geodesic. □

Note in Corollary 4.14 that since the null geodesic is maximal, it cannot contain any null conjugate points to p prior to q [cf. O'Neill (1983, p. 404)]. A sometime useful special case of Corollary 4.14 occurs when $p = q$ is assumed. In this case, one may deduce that if the space–time (M,g) is chronological but not causal, then there exists a smooth null geodesic $\beta : [0,1] \to (M,g)$ with $\beta(0) = \beta(1)$ and $\beta'(1) = \lambda\beta'(0)$ for some $\lambda > 0$. (If $\beta'(1)$ and $\beta'(0)$

were not proportional, then β could be deformed near $\beta(0)$ to be future timelike, contradicting the condition $d(m, m) = 0$ for all m in M, since (M, g) is chronological [cf. Proposition 2.19 of Penrose (1972, p. 15)]. Even more dramatically, a closed timelike curve could be produced to violate chronology [cf. Proposition 10.46 of O'Neill (1983, pp. 294–295)].)

As a second application of the elementary properties of the distance function, we give a proof of the existence of a smooth closed timelike geodesic on any compact space–time having a regular cover with a compact Cauchy surface. Using infinite-dimensional Morse theory, it may be shown [cf. Klingenberg (1978)] that any compact Riemannian manifold admits at least one smooth closed geodesic. However, the method of proof relies crucially on the positive definiteness of the metric and thus is not applicable to Lorentzian manifolds. Nonetheless, one may obtain the following theorem of Tipler by direct methods [cf. Tipler (1979) for a stronger result].

Theorem 4.15. *Let (M, g) be a compact space–time with a regular covering space which is globally hyperbolic and has a compact Cauchy surface. Then (M, g) contains a closed timelike geodesic.*

Proof. Since M is compact, there exists a closed, future directed, timelike curve $\gamma : [0, 1] \to M$. Set $p = \gamma(0) = \gamma(1)$. Let $\pi : \widetilde{M} \to M$ denote the given covering manifold, and let $\widetilde{\gamma} : [0, 1] \to \widetilde{M}$ be a lift of γ, i.e., $\pi \circ \widetilde{\gamma}(t) = \gamma(t)$ for all $t \in [0, 1]$. Then $\widetilde{\gamma}$ is a future directed timelike curve in $\widetilde{M}$. Put $p_1 = \widetilde{\gamma}(0)$ and $p_2 = \widetilde{\gamma}(1)$. Then the global hyperbolicity of $\widetilde{M}$ implies p_1 and p_2 are distinct points which cannot lie on any common Cauchy surface. Since $\pi : \widetilde{M} \to M$ is regular, there must be a deck transformation $\psi : \widetilde{M} \to \widetilde{M}$ taking p_1 to p_2 [cf. Wolf (1974, pp. 35–38, 60)]. Choose a compact Cauchy surface S_1 of $\widetilde{M}$ containing p_1, and define $S_2 = \psi(S_1)$. Since $(\widetilde{M}, \widetilde{g})$ is globally hyperbolic, the distance function $d = d(\widetilde{g}) : \widetilde{M} \times \widetilde{M} \to \mathbb{R} \cup \{\infty\}$ is finite-valued and continuous. Thus we may define a continuous function $f : S_1 \to \mathbb{R}$ by $f(s) = d(s, \psi(s))$. Since $f(p_1) > 0$, we have $A = \sup\{d(s, \psi(s)) : s \in S_1\} > 0$. Moreover, since S_1 is compact, $A < \infty$ and there exists an $r_1 \in S_1$ with $d(r_1, \psi(r_1)) = A$. Let $\widetilde{c} : [0, 1] \to \widetilde{M}$ be a timelike geodesic segment with $\widetilde{c}(0) = r_1$, $\widetilde{c}(1) = \psi(r_1)$, and $L(\widetilde{c}) = d(r_1, \psi(r_1)) = A$. This geodesic exists since $(\widetilde{M}, \widetilde{g})$ is globally hyperbolic. Because $\widetilde{g} = \pi^* g$, it follows that $c = \pi \circ \widetilde{c} : [0,1] \to M$ is a timelike geodesic. Since $\widetilde{c}(0) = r_1$ and $\widetilde{c}(1) = \psi(r_1)$, we also have $c(0) = c(1)$. If c were not smooth at $c(0)$, we could deform c to a timelike curve $\sigma : [0,1] \to M$ with $L_g(\sigma) > L(c)$, $\sigma(0) = \sigma(1) \in \pi(S_1)$, which lifts to a curve $\widetilde{\sigma} : [0,1] \to \widetilde{M}$ with $\widetilde{\sigma}(0) \in S_1$ and $\widetilde{\sigma}(1) = \psi(\widetilde{\sigma}(0)) \in S_2$. But then $L_{\widetilde{g}}(\widetilde{\sigma}) = L_g(\sigma) > L_g(c) = L_{\widetilde{g}}(\widetilde{c}) = A$, in contradiction. □

More recently, G. Galloway (1984b) has obtained, by elementary geometric arguments, the existence of a closed timelike geodesic from any *stable* nontrivial free homotopy class of closed timelike curves on an arbitrary compact space–time. Galloway's approach is based on geodesic convexity methods which were earlier used to obtain a result, basic to the development of global Riemannian geometry during the early part of this century, given first by J. Hadamard for surfaces and then by E. Cartan for general Riemannian manifolds of higher dimension. Their result concerns the existence, within any nontrivial free homotopy class of curves on a compact Riemannian manifold, of a shortest curve in the given homotopy class, which must then be a nontrivial smooth closed geodesic. From a directly geometric viewpoint, certain basic ideas involved in the existence of this geodesic are the following.

Let $L_0 > 0$ denote the infimum of all lengths of curves in this homotopy class. Choose a minimizing sequence $\{c_k\}$ with $\lim L(c_k) = L_0$ in the given free homotopy class. By compactness and convexity radius arguments, one covers the given manifold by a finite number of geodesically convex neighborhoods (each having compact closure in a larger geodesically convex neighborhood). Using these neighborhoods in succession, each c_k may be approximated by a piecewise smooth geodesic, or equivalently, may be viewed as an N-tuple of successive points $p_1(k), p_2(k), \ldots, p_N(k)$, where a uniform bound needs to be given for the number N of points required. Since a geodesic segment in a convex neighborhood is the shortest curve between any two of its points, this approximation procedure produces a shorter curve than c_k, but one which is also still homotopic to c_k, and hence is in the given free homotopy class. By compactness, a diagonalizing argument gives points $p_1, p_2, \ldots, p_N$ which are limits of a subsequence of all of the above sequences. Joining p_i to p_{i+1}

successively produces a piecewise smooth geodesic c in the given free homotopy class which realizes the minimal length L_0. Now the geodesic c must in fact be smooth at the p_i's, or an even shorter closed curve in the free homotopy class could be produced by the usual "rounding off the corners" procedure. Since $L_0 > 0$, this closed geodesic c must be nontrivial, i.e., not a "point curve." A detailed discussion of the steps involved in the above argument may be found in Spivak (1979, p. 358).

Now in carrying these ideas over to the space–time setting, it is clear that "minimal geodesic segment" should be replaced by "maximal timelike geodesic segment." Technical difficulties arise, however, since the set of unit timelike tangent vectors based at a given point is not a compact set. Hence, problems can arise with timelike tangent vectors tending toward a null direction when trying to do subsequence arguments. Equivalently, in the above context, it is necessary to prevent timelike geodesic segments joining $p_i(k)$ to $p_{i+1}(k)$ from converging to a *null* geodesic segment from p_i to p_{i+1} when the diagonalization procedure is carried out. Indeed, Galloway (1986b) gives an example of a compact space–time which contains no closed timelike geodesics but which contains a closed null geodesic.

In view of these difficulties, Galloway (1984b) considers timelike free homotopy classes which are "stable" for a given compact space–time (M, g_0). Here a given free timelike homotopy class $\mathfrak{C}$ for (M, g_0) is said to be stable if there exists a "wider" Lorentzian metric g for M, i.e., $g_0 < g$ in the sense of the discussion following Remark 3.14, such that if L_g denotes the Lorentzian arc length for (M, g), then the given timelike homotopy class $\mathfrak{C}$ satisfies the condition

$$\sup_{c \in \mathfrak{C}} L_g(c) < +\infty.$$

Galloway (1984b) shows that this concept of stability gives the control needed to force convergence arguments to be successful and hence obtains the theorem that for any compact Lorentzian manifold, each stable free timelike homotopy class contains a longest closed timelike curve which is of necessity a closed timelike geodesic. Galloway also shows how his result may be used to recover Tipler's Theorem 4.15 given above and gives a criterion for a free homotopy class of closed timelike curves to be stable. Finally, we note that in Galloway (1986b) it is shown by covering space arguments that every compact two-dimensional Lorentzian manifold contains a closed timelike or null geodesic. Here one uses dimension two in the essential way that a closed timelike curve for (M, g) corresponds to a closed spacelike curve (hence a spacelike hypersurface) for $(M, -g)$, which is also a Lorentzian manifold since $\dim(M) = 2$. Thus, certain techniques in general relativity involving spacelike hypersurfaces may be applied to the existence problem.

4.2 Distance Preserving and Homothetic Maps

Myers and Steenrod (1939) and Palais (1957) have shown that if f is a distance preserving map of a Riemannian manifold (N_1, g_1) onto a Riemannian manifold (N_2, g_2), then f is a diffeomorphism which preserves the metric tensors, i.e., $f^* g_2 = g_1$. In particular, every distance preserving map of (N_1, g_1) onto itself is a smooth isometry. In this section we give similar results for Lorentzian manifolds following Beem (1978a).

Recall that a diffeomorphism $f : (M_1, g_1) \to (M_2, g_2)$ of the Lorentzian manifold (M_1, g_1) onto the Lorentzian manifold (M_2, g_2) is said to be *homothetic* if there exists a constant $c > 0$ such that $g_2(f_* v, f_* w) = c\, g_1(v, w)$ for all $v, w \in T_p M_1$ and all $p \in M_1$. In particular, if $c = 1$, then f is a (smooth) isometry. The group of homothetic transformations is important in general relativity since it has been shown to be the group of transformations which preserves the causal structure for a large class of space–times [cf. Zeeman (1964, 1967), Göbel (1976)].

We will let d_1 denote the Lorentzian distance function of (M_1, g_1) and d_2 denote the Lorentzian distance function of (M_2, g_2) below. The distance analogue of a smooth homothetic map is defined as follows.

Definition 4.16. (*Distance Homothety*) A map $f : (M_1, g_1) \to (M_2, g_2)$ is said to be *distance homothetic* if there exists a constant $c > 0$ such that $d_2(f(p), f(q)) = c\, d_1(p, q)$ for all $p, q \in M$. If $c = 1$, then f is said to be *distance preserving.*

It is important to note that for arbitrary Lorentzian manifolds, distance preserving maps are not necessarily continuous. For if (M, g) is a totally vicious space–time, we have seen that $d(p, q) = \infty$ for all $p, q \in M$ [cf. Lemma 4.2–(2)]. Hence any set theoretic bijection $f : M \to M$ is distance preserving but need not be continuous.

Theorem 4.17. *Let (M_1, g_1) be a strongly causal space–time, and let (M_2, g_2) be an arbitrary space–time. If $f : (M_1, g_1) \to (M_2, g_2)$ is a distance homothetic map (not assumed to be continuous) of M_1 onto M_2, then f is a smooth homothetic map. That is, f is a diffeomorphism, and there exists a constant $c > 0$ such that $f^* g_2 = c\, g_1$. In particular, every map of a strongly causal space–time (M, g) onto itself which preserves Lorentzian distance is an isometry.*

Corollary 4.18. *If (M, g) is a strongly causal space–time, then the space of distance homothetic maps of (M, g) equipped with the compact-open topology is a Lie group.*

Proof of Corollary 4.18. Since (M, g) is strongly causal, this group coincides by Theorem 4.17 with the space of smooth homothetic maps of M onto itself which preserve the time orientation. But this second group is a Lie group. □

The proof of Theorem 4.17 will be broken up into a series of lemmas.

Lemma 4.19. *Let (M_1, g_1) and (M_2, g_2) be space–times, and consider a map $f : (M_1, g_1) \to (M_2, g_2)$ which is onto but not necessarily continuous. If f is distance homothetic, then*

(1) *$p \ll q$ iff $f(p) \ll f(q)$, and*
(2) *$f(I^+(p) \cap I^-(q)) = I^+(f(p)) \cap I^-(f(q))$.*

Proof. First (1) holds since $d_2(f(p), f(q)) = c\, d_1(p, q)$, and $p \ll q$ [respectively, $f(p) \ll f(q)$] iff $d_1(p, q) > 0$ [respectively, $d_2(f(p), f(q)) > 0$]. Since (1) implies $p \ll r \ll q$ iff $f(p) \ll f(r) \ll f(q)$, statement (2) follows. □

The importance of (2) stems from the fact that if (M, g) is strongly causal, the sets $\{I^+(p) \cap I^-(q) : p, q \in M\}$ form a basis for the topology of M. Recall that a map $f : M_1 \to M_2$ is said to be *open* if f maps each open set in M_1 to an open set in M_2.

Lemma 4.20. *Let (M_1, g_1) be strongly causal and let (M_2, g_2) be an arbitrary space–time. If f is a (not necessarily continuous) distance homothetic map of (M_1, g_1) onto (M_2, g_2), then f is open and one-to-one.*

Proof. The openness of f is immediate from part (2) of Lemma 4.19. It remains to show that f is one-to-one. Assume there are distinct points p and q of M_1 with $f(p) = f(q)$. Let $U(p)$ be an open neighborhood of p with $q \notin U(p)$ and such that no nonspacelike curve intersects $U(p)$ more than once. Choose $r_1, r_2 \in U(p)$ with $r_1 \ll p \ll r_2$. Clearly, $q \notin I^+(r_1) \cap I^-(r_2)$. It follows from Lemma 4.19 that $f(r_1) \ll f(p) = f(q) \ll f(r_2)$ implies $r_1 \ll q \ll r_2$. This yields $q \in I^+(r_1) \cap I^-(r_2)$ which is a contradiction. □

Applying Lemma 4.20 to f and f^{-1} we obtain

Proposition 4.21. *Let (M_1, g_1) be strongly causal, and let (M_2, g_2) be an arbitrary space–time. Let f be a (not necessarily continuous) map of M_1 onto M_2. If f is distance homothetic, then f is a homeomorphism and (M_2, g_2) is strongly causal.*

Proof. The relation f^{-1} is a function since f is one-to-one by Lemma 4.20. Furthermore, f^{-1} is continuous since Lemma 4.20 shows f is an open map.

In order to complete the proof it is sufficient to show M_2 is strongly causal since Lemma 4.20 will then imply f^{-1} is an open map, whence f is continuous. Given $p' \in M_2$, let $p = f^{-1}(p')$. If $r' \ll p' \ll q'$, then Lemma 4.19 applied to the distance homothetic map f^{-1} yields $f^{-1}(r') \ll p \ll f^{-1}(q')$. Let $U'(p')$ be an open neighborhood of p'. Choose $V'(p') \subseteq U'(p')$ with the closure of $V'(p')$ a compact set contained in an open convex normal neighborhood $W'(p')$ of p'. We may assume that $(W'(p'), g_2|_{W'(p')})$ is globally hyperbolic. Let $\{r'_n\}$ and $\{q'_n\}$ be sequences in $V'(p')$ such that $r'_n \to p'$, $q'_n \to p'$, and $r'_n \ll p' \ll q'_n$ for all n. Assume the strong causality of M_2 fails at p'. This means that for each n, the set $I^+(r'_n) \cap I^-(q'_n)$ cannot be contained in the convex normal neighborhood $W'(p')$ because otherwise the sets $I^+(r'_n) \cap I^-(q'_n)$ would give arbitrarily small neighborhoods of p' which each nonspacelike curve intersects at most once. Choose a sequence of points $\{z'_n\}$ contained in the boundary of $V'(p')$ with $z'_n \in I^+(r'_n) \cap I^-(q'_n)$ for each n. The sequence $\{z'_n\}$ has an accumulation point z because the closure of $V'(p')$ is compact. Furthermore, $f^{-1}(z'_n) \in f^{-1}(I^+(r'_n) \cap I^-(q'_n)) = I^+(f^{-1}(r'_n)) \cap I^-(f^{-1}(q'_n))$. The continuity of f^{-1} implies that $f^{-1}(r'_n) \to p$ and $f^{-1}(q'_n) \to p$. The strong causality of M_1 yields that the sets $I^+(f^{-1}(r'_n)) \cap I^-(f^{-1}(q'_n))$ are approaching the point p. Thus, $f^{-1}(z'_n) \to p$ which means $f^{-1}(z) = p = f^{-1}(p')$. This contradicts the one-to-one property of f^{-1}. Consequently, M_2 must be strongly causal, and the proposition is established. □

Consider the strongly causal space–time M. Given $p \in M$, let $U(p)$ be a convex normal neighborhood of p. The set $U(p)$ may be chosen so small that whenever $q, r \in U(p)$ with $q \leqslant r$, the distance $d(q, r)$ is the length of the unique geodesic segment $\alpha(q, r)$ from q to r which lies in $U(p)$. Furthermore, $U(p)$ may be chosen such that if $q, z, r \in U(p)$ with $q \ll z \ll r$, then the reverse triangle inequality $d(q, r) \geq d(q, z) + d(z, r)$ is valid with strict equality if and only if z is on the geodesic segment from q to r in $U(p)$. Thus, timelike geodesics in a strongly causal space–time are characterized by the space–time distance function, and it follows that distance homothetic maps take timelike geodesics to timelike geodesics.

Lemma 4.22. *If f is a distance homothetic map defined on a strongly causal space–time, then f maps null geodesics to null geodesics.*

Proof. Let $U(p)$ be a convex normal neighborhood of p as in the above paragraph, chosen sufficiently small such that $f(U(p))$ lies in a convex normal neighborhood of $f(p)$. Let $\alpha(q, r)$ be a null geodesic in $U(p)$. Choose $q_n \to q$ and $r_n \to r$ with $q_n \ll r_n$ for all n. Proposition 4.21 then implies that $f(q_n) \to f(q)$ and $f(r_n) \to f(r)$. The map f takes the timelike geodesic $\alpha(q_n, r_n)$ with endpoints q_n and r_n to the timelike geodesic $\alpha(f(q_n), f(r_n))$. Since the geodesics $\alpha(q_n, r_n)$ converge to $\alpha(q, r)$ and the geodesics $\alpha(f(q_n), f(r_n))$ converge to $\alpha(f(q), f(r))$, it follows that f maps $\alpha(q, r)$ to $\alpha(f(q), f(r))$. □

Proof of Theorem 4.17. The fact that f is a diffeomorphism follows from a result proved by Hawking, King, and McCarthy (1976) which states that a homeomorphism which maps null geodesics to null geodesics must be a diffeomorphism. Since M_1 and M_2 are strongly causal, for each $p \in M_1$ there exists a convex normal neighborhood $U_1(p)$ such that for $q \in U_1(p)$ with $p \ll q$, the lengths of the timelike geodesics $\alpha(p,q)$ joining p to q and $\alpha(f(p), f(q))$ joining $f(p)$ to $f(q)$ are given by $d_1(p,q)$ and $d_2(f(p), f(q))$, respectively. Using $d_2(f(p), f(q)) = c\, d_1(p,q)$, it follows that f maps g_1 onto the tensor $c^{-2}g_2$. □

It is well known that if a complete Riemannian manifold is not locally flat, then it admits no homothetic maps that are not isometries [cf. Kobayashi and Nomizu (1963, p. 242, Lemma 2)]. An essential step in the proof consists of showing for arbitrary complete Riemannian manifolds that any homothetic map which is not an isometry has a unique fixed point. This may be done by using the triangle inequality for the Riemannian distance function and the metric completeness of any geodesically complete Riemannian manifold.

In view of Theorem 4.17 above, it is then of interest to consider the analogous question of the existence of nonisometric homothetic maps of a Lorentzian manifold [cf. Beem (1978b)]. In what follows, we will use the standard terminology of *proper* homothetic map for a homothetic map which is not an isometry.

We first note that $\mathbb{R}^2$ with the Lorentzian metric $ds^2 = dxdy$ provides an example of a globally hyperbolic geodesically complete space–time that admits a fixed-point free, proper homothetic map. For fixing any $\beta \neq 0$ and choosing any $c > 0$, the map $f(x,y) = (x+\beta, cy)$ is a fixed-point free homothetic map with homothetic constant c. Thus the existence of a fixed point for a proper homothetic map must be assumed for geodesically complete Lorentzian manifolds unlike the Riemannian case.

Now suppose f is a proper homothetic map of a space–time (M,g) such that $f(p) = p$ for some $p \in M$. Then $f_{*_p} : T_pM \to T_pM$ has at least one nonspacelike eigenvector [cf. Beem (1978b, p. 319, Lemma 3)]. This eigenvector may be null, however. For example, composing the Lorentzian "boost" isometry F of $(\mathbb{R}^2, ds^2 = dx^2 - dy^2)$,

$$F(x,y) = (x \cosh t + y \sinh t, x \sinh t + y \cosh t)$$

with $t > 0$ fixed, and a dilation $T(x,y) = (cx, cy)$ with $c > 0$ and $c \neq 1$ yields a proper homothetic map f of Minkowski space–time fixing the origin such that $f_{*(0,0)}$ has null vectors for eigenvectors.

But if $f_{*_p} : T_pM \to T_pM$ is a proper homothetic map which has a timelike eigenvector with eigenvalue $\lambda < 1$, it may be shown that (M, g) is Minkowski space–time [cf. Beem (1978b, p. 319, Proposition 4)]. Also, if f is a homothetic map with a fixed point p such that all eigenvalues of f_{*_p} are real and all have absolute value less than one, then (M, g) is Minkowski space–time [cf. Beem (1978b, p. 316, Theorem 1)].

We now give an example of a nonflat space–time admitting a global homothetic flow. Let $M = \mathbb{R}^3$ with the metric $g = ds^2 = e^{xz}dxdy + dz^2$. Thus if

$$v = a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y} + c\frac{\partial}{\partial z}, \qquad w = \overline{a}\frac{\partial}{\partial x} + \overline{b}\frac{\partial}{\partial y} + \overline{c}\frac{\partial}{\partial z}$$

are tangent vectors at (x, y, z), we have

$$g(v,w) = e^{xz}\frac{\overline{a}b + a\overline{b}}{2} + c\overline{c}.$$

It may then be checked that while (M, g) is not flat, the map $\phi_t : (\mathbb{R}^3, ds^2) \to (\mathbb{R}^3, ds^2)$ given by

$$\phi_t(x,y,z) = (e^t x, e^{-3t} y, e^{-t} z)$$

is a proper homothety with $g(\phi_{t_*} v, \phi_{t_*} w) = e^{-2t} g(v,w)$ for each fixed nonzero t.

We now show, however, that this space–time is null geodesically incomplete. Let $X = \partial/\partial x$, $Y = \partial/\partial y$, and $Z = \partial/\partial z$. Then all inner products vanish except for $g(X,Y) = e^{xz}/2$ and $g(Z,Z) = 1$; furthermore, $[X,Y] = [X,Z] = 0$. Hence using the Koszul formula

$$\begin{aligned} 2\,g(\nabla_U V, W) &= Ug(V,W) + Vg(U,W) - Wg(U,V) \\ &\quad + g([U,V],W) - g([U,W],V) - g([V,W],U), \end{aligned}$$

we obtain the following formulas for the Levi–Civita connection of $(\mathbb{R}^3, ds^2)$:

$$\nabla_X X = zX, \qquad \nabla_Y Y = \nabla_Z Z = 0,$$

$$\nabla_X Y = -\frac{x}{4}e^{xz}Z, \qquad \nabla_X Z = \frac{x}{2}X, \qquad \text{and} \qquad \nabla_Y Z = \frac{x}{2}Y.$$

Thus the only nonzero Christoffel symbols are $\Gamma_{11}^1 = z$, $\Gamma_{12}^3 = \Gamma_{21}^3 = -\frac{x}{4}e^{xz}$, and $\Gamma_{13}^1 = \Gamma_{31}^1 = \Gamma_{23}^2 = \Gamma_{32}^2 = \frac{x}{2}$. Hence if $\gamma(t) = (x(t), y(t), z(t))$ is a geodesic, the usual system of second order differential equations

$$\frac{d^2x_k}{dt^2} + \sum_{i,j} \Gamma_{ij}^k \circ \gamma(t) \frac{dx_i}{dt}\frac{dx_j}{dt} = 0$$

for γ reduces to the following system:

$$\begin{aligned} x'' + z(x')^2 + xx'z' &= 0, \\ y'' + xy'z' &= 0, \\ z'' - \frac{e^{xz}}{2}xx'y' &= 0. \end{aligned}$$

It may finally be checked that the curve $\gamma : (-1, \infty) \to (\mathbb{R}^3, ds^2)$ defined by $\gamma(t) = (\ln(1+t), 0, 1)$ satisfies this system of differential equations and hence is the unique null geodesic in $(\mathbb{R}^3, ds^2)$ with $\gamma'(0) = \partial/\partial x|_{(0,0,1)}$. Thus this space–time is null geodesically incomplete.

4.3 The Lorentzian Distance Function and Causality

In this section we study the relationship between the continuity and finiteness of the Lorentzian distance function $d = d(g) : M \times M \to \mathbb{R} \cup \{\infty\}$ for (M, g) and the causal structure of (M, g). The most elementary properties, extending Lemma 4.2 above, are summarized in the following lemma. Recall that $\text{Lor}(M)$ denotes the space of all Lorentzian metrics for M. The C^0 topology on $\text{Lor}(M)$ was defined in Section 3.2.

Lemma 4.23.

(1) $d(p, q) > 0$ *iff* $q \in I^+(p)$.

(2) *The space–time* (M, g) *is totally vicious iff* $d(p, q) = \infty$ *for all* $p, q \in M$.

(3) *The space–time (M, g) is chronological iff d is identically zero on the diagonal $\Delta(M) = \{(p,p) : p \in M\}$ of $M \times M$.*

(4) *The space–time (M, g) is future [respectively, past] distinguishing iff for each pair of distinct $p, q \in M$, there is some $x \in M$ such that exactly one of $d(p,x)$ and $d(q,x)$ [respectively, $d(x,p)$ and $d(x,q)$] is zero.*

(5) *The space–time (M, g) is stably causal iff there exists a neighborhood U of g in the fine C^0 topology on $\mathrm{Lor}(M)$ such that $d(g')(p,p) = 0$ for all $g' \in U$ and $p \in M$.*

Proof. Similar to Lemma 4.2 and Remark 4.3. □

Recall that the Lorentzian distance function in general fails to be upper semicontinuous. Thus the continuity of $d(g)$ should have implications for the causal structure of (M, g). An example is the following result first stated in Beem and Ehrlich (1977, p. 1130). Here d is regarded as being continuous at $(p,q) \in M \times M$ with $d(p,q) = \infty$ because $d(p_n, q_n) \to \infty$ for all sequences $p_n \to p$ and $q_n \to q$ (cf. Lemma 4.4).

Theorem 4.24. *Let (M, g) be a distinguishing space–time. If $d = d(g) : M \times M \to \mathbb{R} \cup \{\infty\}$ is continuous, then (M, g) is causally continuous.*

Proof. We need only show that I^+ and I^- are outer continuous. Assume I^+ is not outer continuous. There is then some compact set $K \subseteq M - \overline{I^+(p)}$ and some sequence $p_n \to p$ such that $K \cap \overline{I^+(p_n)} \neq \emptyset$ for all n. Let $q_n \in K \cap \overline{I^+(p_n)}$ and let $\{q_m\}$ be a subsequence of $\{q_n\}$ such that $\{q_m\}$ converges to some point q of the compact set K. Then $q_m \to q$ and $q_m \in \overline{I^+(p_m)}$ imply there must be a sequence $\{q'_m\}$ converging to q such that $q'_m \in I^+(p_m)$ for each m. Since $M - \overline{I^+(p)}$ is an open neighborhood of q, there is some $r \in M - I^+(p)$ with $q \ll r$. For sufficiently large m we then have $q'_m \ll r$ and hence $p_m \ll q'_m \ll r$. Thus $d(p_m, r) \geq d(p_m, q'_m) + d(q'_m, r)$. Using the lower semicontinuity of distance and the causality relation $q \ll r$, we obtain $0 < d(q,r) \leq \liminf d(q'_m, r)$. Consequently, $d(p_m, r) \geq d(q,r)/2 > 0$ for all sufficiently large m. However, since $r \notin I^+(p)$, we have $d(p,r) = 0$, and hence $d(p,r) \neq \lim d(p_m, r)$. Thus if d is continuous, I^+ is outer continuous.

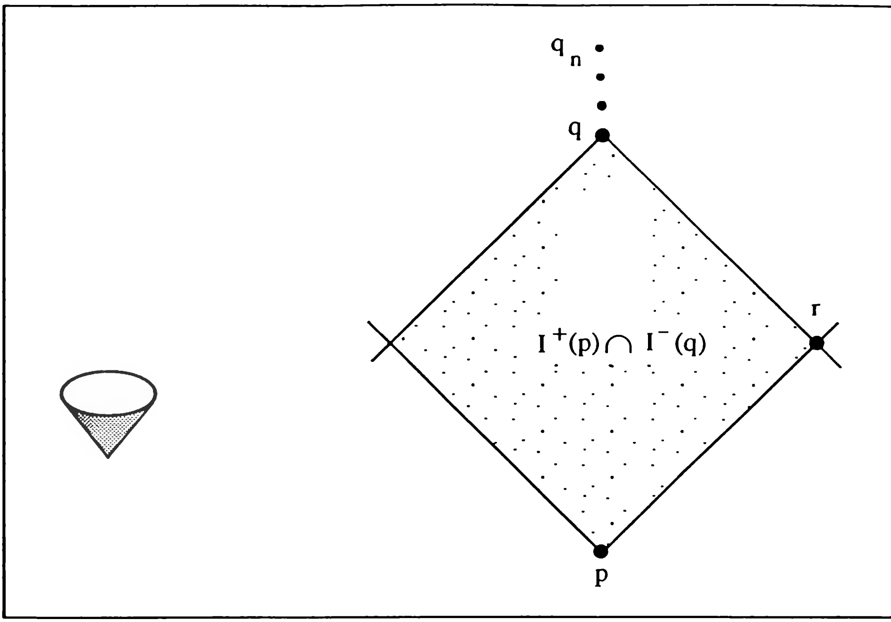


FIGURE 4.6. Let (M, g) denote Minkowski space–time with the point r deleted. Choose $p, q \in M$ such that in Minkowski space–time the point r is on the boundary of $I^+(p) \cap I^-(q)$ as shown. Let $\{q_n\}$ be a sequence of points approaching q with $q \ll q_n$ for each n. There is a smooth conformal factor $\Omega : M \to (0, \infty)$ such that $\Omega \equiv 1$ on $I^+(p) \cap I^-(q)$ and yet $d(\Omega g)(p, q_n) \geq 2d(g)(p, q)$ for each n. The function Ω will be unbounded near the deleted point r. Since $d(g)(p, q) = d(\Omega g)(p, q) < \liminf d(\Omega g)(p, q_n)$, the causally continuous space–time $(M, \Omega g)$ has a Lorentzian distance function which is discontinuous at $(p, q) \in M \times M$.

A similar argument shows that I^- is outer continuous. Thus continuity of d implies that (M, g) is causally continuous. □

The following example shows that the converse of Theorem 4.24 is false. Let (M, g) denote Minkowski space–time with a single point removed. The space–time $(M, \Omega g)$ will be causally continuous for any smooth conformal factor $\Omega : M \to (0, \infty)$. However, Ω may be chosen such that $d = d(\Omega g)$ is *not* continuous (cf. Figure 4.6).

We now turn to a characterization of strongly causal space–times in terms of the Lorentzian distance function. The definition of convex normal neighborhood was given in Section 3.1. Given any space–time (M, g), we have

Definition 4.25. (*Local Distance Function*) A *local distance function* (D, U) on (M, g) is a convex normal neighborhood U together with the distance function $D : U \times U \to \mathbb{R}$ induced on U by the space–time $(U, g|_U)$.

More explicitly, given $p, q \in U$, then $D(p, q) = 0$ if there is no future directed timelike geodesic segment in U from p to q. Otherwise, $D(p, q)$ is the Lorentzian arc length of the unique future directed timelike geodesic segment in U from p to q.

We will let $I^+(p, U)$ (respectively, $J^+(p, U)$) denote the chronological (respectively, causal) future of p with respect to the space–time $(U, g|_U)$.

Lemma 4.26. *Let (M, g) be a space–time, and let U be a convex normal neighborhood of (M, g). Assume that $D : U \times U \to \mathbb{R}$ is the distance function for $(U, g|_U)$. Then D is a continuous function on $U \times U$ and D is differentiable on $U^+ = \{(p, q) \in U \times U : q \in I^+(p, U)\}$.*

Proof. Given $p, q \in U$ with $q \in J^+(p, U)$, let $c_{pq} : [0, 1] \to U$ denote the unique nonspacelike geodesic segment with $c_{pq}(0) = p$ and $c_{pq}(1) = q$. We then have $D(p, q) = [-g(c'_{pq}(0), c'_{pq}(0))]^{1/2}$ and $D(p, q)^2 = [-g(c'_{pq}(0), c'_{pq}(0))]$. From the differentiable dependence of geodesics on endpoints in convex neighborhoods, it is immediate that D is continuous on $U \times U$ and that D is differentiable on U^+. □

Minkowski space–time shows that D fails to be differentiable across the null cones and thus fails to be smooth on all of $U \times U$. It is not hard to see that the local distance function (D, U) uniquely determines the Lorentzian metric g on U. Consequently if $\{U_\alpha\}$ is a covering of M by convex normal neighborhoods with associated local distance functions $\{(D_\alpha, U_\alpha)\}$, then $\{(D_\alpha, U_\alpha)\}$ uniquely determines g on M.

We now characterize strongly causal space–times in terms of local distance functions [cf. Beem and Ehrlich (1979c, Theorem 3.4)].

Theorem 4.27. *A space–time (M, g) is strongly causal if and only if each point $r \in M$ has a convex normal neighborhood U such that the local distance function (D, U) agrees on $U \times U$ with the distance function $d = d(g) : M \times M \to \mathbb{R} \cup \{\infty\}$.*

Proof. If (M, g) is strongly causal and $r \in M$, then there is some convex normal neighborhood U of r such that no nonspacelike curve which leaves U ever returns. The local distance function for U then agrees with $d = d(g) \mid (U \times U)$.

Conversely, assume that strong causality breaks down at some point $r \in M$. Let U be a convex normal neighborhood of r such that $D(p, q) = d(p, q)$ for all $p, q \in U$. There exists a neighborhood $W \subseteq U$ of r such that any future directed nonspacelike curve $\gamma : (0, 1] \to U$ with $\gamma(1) \in W$ and γ past inextendible in U contains some point not in $J^+(W, U)$. Since strong causality fails to hold at r, there is a future directed timelike curve $\gamma_1 : [0, 1] \to M$ with $r' = \gamma_1(0) \in W$, $\gamma_1(1/2) \notin U$, and $\gamma_1(1) \in W$. By construction of W, there is some point $p \in \gamma_1 \cap U$ with $p \notin J^+(r', U)$. Hence $D(r', p) = 0$. However, $d(r', p) > 0$ since $d(r', p)$ is at least as large as the length of γ_1 from r' to p. Thus $D(r', p) \neq d(r', p)$. Taking the contrapositive then establishes the theorem. □

Corollary 4.28. *If (M, g) is strongly causal, then d is continuous on some neighborhood of $\Delta(M) = \{(p, p) : p \in M\}$ in $M \times M$. Also, given any point $m \in M$, there exists a convex normal neighborhood U of m such that $d \mid (U \times U)$ is finite-valued.*

We now give a characterization of globally hyperbolic space–times among all strongly causal space–times using the Lorentzian distance function. For this purpose, it is first necessary to show that the usual definition of globally hyperbolic may be weakened.

Lemma 4.29. *Let (M, g) be a strongly causal space–time. If $J^+(p) \cap J^-(q)$ has compact closure for all $p, q \in M$, then (M, g) is globally hyperbolic.*

Proof. It is only necessary to show $J^+(p) \cap J^-(q)$ is always closed. Assume $r \in (\overline{J^+(p) \cap J^-(q)}) - (J^+(p) \cap J^-(q))$. Choose a sequence $\{r_n\}$ of points in $J^+(p) \cap J^-(q)$ with $r_n \to r$. For each n let $\gamma_n : [0, 1) \to M$ be a future directed, future inextendible, nonspacelike curve with $p = \gamma_n(0)$ and $r_n \in \gamma_n$, $q \in \gamma_n$. By Proposition 3.31, there is some future directed, future inextendible, nonspacelike limit curve $\gamma : [0, 1) \to M$ of the sequence $\{\gamma_n\}$. Furthermore, $p = \gamma(0)$. The limit curve γ cannot be future imprisoned in any compact subset of M because (M, g) is strongly causal (cf. Proposition 3.13). Consequently, there is some point x on γ with $x \notin \overline{J^+(p) \cap J^-(q)}$. The definition of limit curve yields a subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$ and points $x_m \in \gamma_m$ with $x_m \to x$. Since $x \notin \overline{J^+(p) \cap J^-(q)}$, we have $x_m \notin J^+(p) \cap J^-(q)$ for all large m. Using $\gamma_m \subseteq J^+(p)$, it follows that $x_m \notin J^-(q)$ for large m. Hence q lies between p and x_m on γ_m for large m. Let $\gamma[p, x]$ (respectively, $\gamma_m[p, x_m]$) denote the portion of γ (respectively, γ_m) from p to x (respectively, x_m). By Proposition 3.34 we may assume, by taking a subsequence of $\{\gamma_m[p, x_m]\}$ if necessary, that $\{\gamma_m[p, x_m]\}$ converges to $\gamma[p, x]$ in the C^0 topology on curves. Hence $q \in \gamma_m[p, x_m]$ for large m implies $q \in \gamma[p, x]$. Also $r_m \to r$ and $r_m \leqslant q$ which yield $r \in \gamma[p, q]$. Thus $r \in J^+(p) \cap J^-(q)$, in contradiction. □

Recall from Definition 4.6 above that a space–time (M, g) is said to satisfy the *finite distance condition* if and only if $d(g)(p, q) < \infty$ for all $p, q \in M$. This condition may be used to characterize globally hyperbolic space–times among strongly causal space–times [cf. Beem and Ehrlich (1979b, Theorem 3.5)].

Theorem 4.30. *The strongly causal space–time (M, g) is globally hyperbolic iff (M, g') satisfies the finite distance condition for all $g' \in C(M, g)$.*

Proof. It has already been remarked that if (M, g) is globally hyperbolic, then all metrics in $C(M, g)$ satisfy the finite distance condition (cf. Corollary 4.7).

Conversely, assume that (M, g) is not globally hyperbolic. Lemma 4.29 implies that there exist $p, q \in M$ such that $J^+(p) \cap J^-(q)$ does not have compact closure. Let h be an auxiliary geodesically complete positive definite metric on M, and let $d_0 : M \times M \to \mathbb{R}$ be the Riemannian distance function induced on M by h. The Hopf–Rinow Theorem implies that all subsets of M which are bounded with respect to d_0 have compact closure. Thus $J^+(p) \cap J^-(q)$ is not bounded. Hence, for each n we may choose $p_n \in J^+(p) \cap J^-(q)$ such that $d_0(p, p_n) > n$. Choose p' and q' with $p' \ll p \ll q \ll q'$. We wish to show there exists a conformal factor Ω such that $d(\Omega g)(p', q') = \infty$. For each $n > 1$, choose γ_n to be a future directed timelike curve from p' to p_n such that $\gamma_n[1/2, 3/4] \subseteq \{r \in M : n-1 < d_0(p,r) < n\}$. For each $n > 1$, let $\Omega_n : M \to \mathbb{R}$ be a smooth function such that $\Omega_n(x) = 1$ if $x \notin \{r : n-1 < d_0(p,r) < n\}$ and such that the length of $\gamma_n[1/2, 3/4]$ is greater than n for the metric $\Omega_n g$. Let $\Omega = \prod_n \Omega_n$. This infinite product is well defined on M since for each $x \in M$ at most one of the factors Ω_n is not unity. Then $d(\Omega g)(p', p_n) > n$ for each $n > 1$. Hence $d(\Omega g)(p', q') = \infty$ as $d(\Omega g)(p', q') \geq d(\Omega g)(p', p_n) + d(\Omega g)(p_n, q')$ for each n. $\square$

We now turn to the proof that for distinguishing space–times with continuous distance functions, the future and past inner balls form a subbasis for the given manifold topology. Recall that

$$B^+(p, \epsilon) = \{q \in I^+(p) : d(p,q) < \epsilon\} = \{q \in M : 0 < d(p,q) < \epsilon\},$$

and

$$B^-(p, \epsilon) = \{q \in I^-(p) : d(q,p) < \epsilon\} = \{q \in M : 0 < d(q,p) < \epsilon\}.$$

Thus defining $f_i : M \to \mathbb{R}$ for $i = 1, 2$ by $f_1(q) = d(p,q)$ and $f_2(q) = d(q,p)$, we have $B^+(p, \epsilon) = f_1^{-1}((0, \epsilon))$ and $B^-(p, \epsilon) = f_2^{-1}((0, \epsilon))$. Hence if (M, g) has a continuous distance function, the inner balls $B^+(p, \epsilon)$ and $B^-(p, \epsilon)$ of M are open in the manifold topology.

Proposition 4.31. *Let (M, g) be a distinguishing space–time with a continuous distance function. Then the collection*

$$\{B^+(p, \epsilon_1) \cap B^-(q, \epsilon_2) : p, q \in M,\ \epsilon_1, \epsilon_2 > 0\}$$

forms a basis for the given manifold topology of M.

Proof. The above arguments show that sets of the form $B^+(p,\epsilon_1) \cap B^-(q,\epsilon_2)$ are open in the manifold topology. Thus given an arbitrary point $r \in M$ and an arbitrary open neighborhood $U(r)$ of r in the manifold topology, it is sufficient to show that there exist $p, q \in M$ and $\epsilon_1, \epsilon_2 > 0$ with $r \in B^+(p,\epsilon_1) \cap B^-(q,\epsilon_2) \subseteq U(r)$.

Theorem 4.24 yields that (M,g) is causally continuous and hence also strongly causal. Thus we may choose a convex normal neighborhood V of r with $V \subseteq U(r)$ such that no nonspacelike curve which leaves V ever returns and such that $d : V \times V \to \mathbb{R} \cup \{\infty\}$ is finite-valued (cf. Corollary 4.28). Fix $p, q \in V$ with $p \ll r \ll q$. Then $r \in I^+(p) \cap I^-(q) \subseteq V$ since no nonspacelike curve from p to q can leave V and return. Letting $\epsilon_1 = d(p,r) + 1$ and $\epsilon_2 = d(r,q) + 1$, we obtain

$$r \in B^+(p,\epsilon_1) \cap B^-(q,\epsilon_2) \subseteq I^+(p) \cap I^-(q) \subseteq V \subseteq U(r)$$

which establishes the proposition. $\square$

We conclude this section with a characterization of totally geodesic timelike submanifolds in terms of the Lorentzian distance function. An analogous result holds for submanifolds of (not necessarily complete) Riemannian manifolds [cf. Gromoll, Klingenberg, and Meyer (1975, p. 159)].

Let (M,g) denote an arbitrary strongly causal space–time. Suppose that $i : N \to M$ is a smooth submanifold and set $\overline{g} = i^*g$. Recall that $(N, \overline{g})$ is said to be a *timelike submanifold* of (M,g) if $\overline{g}|_p : T_pN \times T_pN \to \mathbb{R}$ is a Lorentzian metric for each $p \in N$. As usual, we will identify N and $i(N)$. Let $\overline{L}$, L and $\overline{d}$, d denote the arc length functionals and Lorentzian distance functions of $(N,\overline{g})$ and (M,g), respectively. Then if γ is a smooth curve in $(N,\overline{g})$, we have $\overline{L}(\gamma) = L(\gamma)$. Note also that if $q \in I^+(p,N)$, then $p \ll q$ in (M,g), and if $q \in J^+(p,N)$, then $p \leqslant q$ in (M,g). Thus it follows immediately from the definitions of d and $\overline{d}$ that

$$\overline{d}(m,n) \leq d(m,n) \qquad \text{for all } m,n \in N. \tag{4.4}$$

With this remark in hand, we are ready to prove the following result.

Proposition 4.32. *Let $(N, \overline{g})$ be a totally geodesic timelike submanifold of the strongly causal space–time (M, g). Then given any $p \in N$, there exists a neighborhood V of p in N such that $d| \, (V \times V) = \overline{d}\,| \, (V \times V)$.*

Proof. First, let W be a convex neighborhood of p in (M, g) such that every pair of points $m, n \in W$ are joined by a unique geodesic of (M, g) lying in W and if $m \leqslant n$, then this geodesic is maximal in (M, g). We may then choose a smaller neighborhood V_0 of p in M with $V_0 \subseteq W$ such that if $V = V_0 \cap N$, then V is contained in a convex normal neighborhood U of p in N with U contained in W.

Suppose first that $m, n \in V$ and $n \in J^+(m, N)$. Since $V \subseteq U$, there exists a nonspacelike geodesic γ of $(N, \overline{g})$ in U from m to n. Also, as N is totally geodesic, γ is a nonspacelike geodesic in (M, g). Because γ is contained in U and $U \subseteq W$, γ is maximal in (M, g). We thus have $\overline{d}(m, n) \geq \overline{L}(\gamma) = L(\gamma) = d(m, n)$. In view of (4.4), we obtain $\overline{d}(m, n) = d(m, n)$ as required.

It remains to consider the case that $m, n \in V$ and $n \notin J^+(m, N)$. Thus $\overline{d}(m, n) = 0$ by definition. Suppose that $d(m, n) > 0$. Then there exists a timelike geodesic γ_1 of (M, g) in W from m to n. On the other hand, since $m, n \in U$, there exists a geodesic γ_2 of $(N, \overline{g})$ from m to n lying in U which must be spacelike since $n \notin J^+(m, N)$. Since $(N, \overline{g})$ is totally geodesic, γ_2 is also a spacelike geodesic of (M, g) from m to n which lies in $U \subseteq W$. Thus we have distinct geodesics γ_1 and γ_2 in W from m to n, in contradiction. Hence $d(m, n) = 0 = \overline{d}(m, n)$ as required. □

We now prove the converse of Proposition 4.32.

Proposition 4.33. *Let $(N, \overline{g})$ be a timelike submanifold of the strongly causal space–time (M, g). Suppose that for all $p \in N$ there exists a neighborhood V of p in N such that $d| \, (V \times V) = \overline{d}\,| \, (V \times V)$. Then $(N, \overline{g})$ is totally geodesic in (M, g).*

Proof. It suffices to fix any $p \in N$ and show that the second fundamental form S_n vanishes at p (cf. Definition 3.48). Since any tangent vector in T_pN may be written as a sum of nonspacelike tangent vectors, it is enough to show that $S_n(v, w) = 0$ for all nonspacelike tangent vectors in T_pN. Also, as

$S_n(-v, w) = -S_n(v, w)$, it suffices to show that $S_n(v, w) = 0$ for all future directed nonspacelike tangent vectors in T_pN.

Thus let $v \in T_pN$ be a future directed nonspacelike tangent vector. Let γ denote the unique geodesic in $(N, \overline{g})$ with $\gamma'(0) = v$. Also let V be a neighborhood of p in N on which the distance functions $\overline{d}$ and d coincide. Choose $t > 0$ such that if $m = \gamma(t)$, then $m \in V$ and $\overline{d}(p, m) = \overline{L}(\gamma \,|\, [0, t])$ is finite. We then obtain

$$d(p, m) \geq L(\gamma \,|\, [0, t]) = \overline{L}(\gamma \,|\, [0, t]) = \overline{d}(p, m).$$

But since $m \in V$ we have $d(p, m) = \overline{d}(p, m)$, whence $L(\gamma \,|\, [0, t]) = d(p, m)$. Hence $\gamma \,|\, [0, t]$ is a geodesic in (M, g) by Theorem 4.13. Thus we have shown that if $v \in T_pN$ is any future directed tangent vector, the geodesic in (M, g) with initial direction v is also a geodesic in $(N, \overline{g})$ near p. Therefore $S_n(v, v) = 0$ for all future directed nonspacelike tangent vectors. Since the sum of two nonparallel future directed nonspacelike tangent vectors is future timelike, it follows by polarization that $S_n(v, w) = 0$ for all future directed nonspacelike tangent vectors $v, w \in T_pN$, as required. $\square$

Combining Propositions 4.32 and 4.33 yields the following characterization of totally geodesic timelike submanifolds of strongly causal space–times in terms of the Lorentzian distance function.

Theorem 4.34. *Let (M, g) be a strongly causal space–time of dimension $n \geq 2$, and suppose that (N, i^*g) is a smooth timelike submanifold of (M, g), i.e., $\overline{g} = i^*g$ is a Lorentzian metric for N. Then $(N, \overline{g})$ is totally geodesic iff given any $p \in N$, there exists a neighborhood V of p in N such that the Lorentzian distance functions $\overline{d}$ of $(N, \overline{g})$ and d of (M, g) agree on $V \times V$.*

4.4 Maximal Geodesic Segments and Local Causality

We have seen in this chapter that certain causality conditions placed on a space–time are related to pleasant local or global behavior of the Lorentzian distance function. For instance, we saw that if (M, g) is globally hyperbolic, then all metrics conformal to g for M are not only continuous but also satisfy the finite distance condition as well. Further, the finite distance condition for all metrics conformal to a given strongly causal metric for a space–time forces the conformal class of space–times to be globally hyperbolic. If a space–time is strongly causal, then we noted that for any given p in M there exists a neighborhood U of p in which the local distance function of $(U, g|_U)$ coincides with the global Lorentzian distance function on $U \times U$. Hence the Lorentzian distance function is forced to be finite-valued on $U \times U$.

For the purposes of this section, it will be convenient to reformulate Definition 4.10 slightly as follows.

Definition 4.35. (*Maximal Segment*) A future directed nonspacelike curve $c : [a, b] \to (M, g)$ is said to be a *maximal segment* provided that $L(c) = d(c(a), c(b))$ and hence $L(c\,|\,[s,t\,]) = d(c(s), c(t))$ for all s, t with $a \leq s \leq t \leq b$.

Evidently, the Lorentzian distance function restricted to the image of a maximal segment must be finite-valued and continuous. In the terminology of Definition 4.35, the previous Proposition 4.12 may be rephrased as follows. Suppose that (M, g) is strongly causal. Given p in M, let U be a local causality neighborhood of p, i.e., a causally convex neighborhood about p which is also a geodesically convex normal neighborhood. Then any nonspacelike geodesic segment lying in U is a maximal segment in the space–time (M, g).

In Chapter 14 the Lorentzian Splitting Theorem for timelike geodesically complete (but not necessarily globally hyperbolic) space–times is studied. Since global hyperbolicity is *not* assumed and hence the global continuity and finite-valuedness of the space–time distance function may not be taken for granted, it is necessary to make a careful study of the space–time distance function and asymptotic geodesics in a neighborhood of a given timelike line. What emerged in a series of papers, as summarized in the introduction to Chapter 14 which will not be repeated here, is that the existence of a global maximal timelike geodesic has important implications for the distance function and the Busemann function of the timelike line in a neighborhood of the given line [cf. Eschenburg (1988), Galloway (1989a), Newman (1990), Galloway and Horta (1995)]. Especially, Newman (1990) made a thorough study of the geodesic geometry in the case that timelike geodesic completeness, but not global hyperbolicity, is assumed. In this section we give certain elementary preliminaries which will be germane to Chapter 14 but which fit into the spirit of this chapter. It is interesting that the reverse triangle inequality plays an important role, hence this material is decidedly non-Riemannian. Also the lower semicontinuity of the Lorentzian distance function for an arbitrary space–time (cf. Lemma 4.4) is useful here.

Note as an immediate first example that if a space–time (M,g) contains a single maximal segment, then (M,g) cannot be totally vicious, since the Lorentzian distance function is finite-valued on the particular segment while totally vicious space–times satisfy $d(p,q) = +\infty$ for all p, q in M. Newman (1990) noted the more interesting consequence that the existence of a maximal timelike segment $c : [a,b] \to (M,g)$ implies that strong causality holds at all points of $c((a,b))$. Since strong causality is an open condition [cf. Penrose (1972, p. 30)], this thus yields an open neighborhood U of $c((a,b))$ for which strong causality at q is valid for all q in U. Hence, not only does the existence of a local causality neighborhood in a strongly causal space–time guarantee the local existence of maximal segments, but a kind of converse holds: the existence of a maximal timelike segment implies that some local region of (M,g) containing all interior points of the given maximal timelike segment must be strongly causal.

For our use in Chapter 14, these basic consequences of the existence of maximal segments will be treated in the present section. We begin with a maximal null segment.

Lemma 4.36. *Let $c : [0,1] \to (M,g)$ be a maximal null geodesic segment, and let $I(c)$ denote the domain of c extended to be a future inextendible null geodesic emanating from $c(0)$. Then*

(1) *For any s, t with $0 \le s < t \le 1$ and $r \in J^+(c(s)) \cap J^-(c(t))$, we have*

$$d(c(s),r) = d(r,c(t)) = 0, \tag{4.5}$$

hence r lies on $c(I(c))$;

(2) *For any p, q in $J^+(c(s)) \cap J^-(c(t))$ with $p \leqslant q$, we have $d(p,q) = 0$; and*

(3) *Chronology holds at all points of $J^+(c(0)) \cap J^-(c(1))$.*

Proof. (1) By assumption, $c(s) \leqslant r \leqslant c(t)$ and $d(c(s), c(t)) = 0$, so that the reverse triangle inequality $d(c(s), c(t)) \geq d(c(s), r) + d(r, c(t))$ implies equation (4.5). Further, since $c(s) \leqslant r \leqslant c(t)$ is assumed, there exist future causal curves c_1 from $c(s)$ to r and c_2 from r to $c(t)$ which by (4.5) must both be maximal null geodesic segments. Since $c(s)$ and $c(t)$ are not chronologically related, the concatenation of c_1 and c_2 must constitute a single null geodesic by basic causality theory [cf. Penrose (1972, Proposition 2.19)], whence $r \in c(I(c))$.
(2) This is immediate from the reverse triangle inequality applied to

$$c(s) \leqslant p \leqslant q \leqslant c(t).$$

(3) Condition (3) now follows since if chronology fails to hold, then $d(p,p)$ is infinite which contradicts (2) applied with $q = p$. □

We should caution that $I(c)$ is not necessarily equal to $[0, +\infty)$ (cf. Lemma 7.4). In the case of a maximal timelike segment, the reverse triangle inequality yields the finiteness of Lorentzian distance from points in some neighborhood of the segment despite the possible general lack of finiteness of distance for chronologically related points (cf. Figure 4.2).

Similar arguments to those used in Lemma 4.36 yield the following finiteness of the distance function in the causal hull of any causal maximal segment.

Lemma 4.37. *Let $c : [0,1] \to (M,g)$ be a causal maximal segment. Then*

(1) *Given any p, q in $J^+(c(0)) \cap J^-(c(1))$ with $p \leqslant q$, the distance $d(p,q)$ is finite.*

(2) *Chronology holds at all points of $J^+(c(0)) \cap J^-(c(1))$.*

Having dealt with chronology, let us now consider the stronger requirement of causality. The cylinder $M = S^1 \times \mathbb{R}$ with metric $ds^2 = d\theta\, dt$ contains closed null geodesics $c : [0, +\infty) \to M$ which satisfy $d(c(0), c(t)) = 0$ for all $t \geq 0$. Hence, the existence of a maximal null geodesic does not imply more than local chronology.

Lemma 4.38. *Let $c : [0,1] \to (M,g)$ be a maximal timelike segment. Then causality holds at all points of $c([0,1])$.*

Proof. Suppose that c_1 is a closed nonspacelike curve beginning and ending at $c(t_1)$ with $t_1 > 0$. Since chronology holds at $c(t_1)$ by Lemma 4.37, c_1 must be a maximal null segment (cf. Corollary 4.14). Consider the composite causal curve $\gamma = (c\,|\,[0,t_1]) * c_1$ from $c(0)$ to $c(t_1)$. Since γ is a causal curve from $c(0)$ to $c(t_1)$ which is a timelike followed by a null geodesic, first variation "rounding the corner" arguments produce a causal curve γ_1 from $c(0)$ to $c(t_1)$ which is longer than γ, hence longer than $c\,|\,[0,t_1]$ [cf. O'Neill (1983, p. 294)]. But this contradicts the maximality of $c\,|\,[0,t_1]$.

If $t_1 = 0$, apply the same type of argument to the composition of the closed null geodesic c_1 followed by $c\,|\,[0,1]$. □

Now we turn to a result with a somewhat more difficult proof first given in Newman (1990, p. 166) and an alternate proof suggested in Galloway and Horta (1995). In the course of the proof, it will be helpful to employ a characterization of the failure of strong causality given by Kronheimer and Penrose [cf. Penrose (1972, p. 31)].

Lemma 4.39. *Let $p \in (M,g)$. Then strong causality fails at p if and only if there exists a point $q \in J^-(p)$ with $q \neq p$ (which may be chosen arbitrarily closely to p) such that $x \ll p$ and $q \ll y$ together imply $x \ll y$ for all x, y in (M,g).*

Proposition 4.40 [Newman (1990)]. *Let $c : [0,a] \to (M,g)$ be a maximal timelike geodesic segment. Then for any t_0 with $0 < t_0 < a$, strong causality holds at $p = c(t_0)$.*

Proof. For convenience, we will assume that c is parametrized by unit speed and also, given $p = c(t_0)$, take $q \in J^-(p)$ in Lemma 4.39 sufficiently close to p that $q \in I^+(c(0))$. Now assume that strong causality fails to hold at p.

We first need to establish that $d(q,p) = 0$, whence q is the initial point of a maximal null segment from q to p. Suppose that $d(q,p) > 0$. Choose a sequence $\{y_n\} \subseteq I^+(q)$ with $y_n \to q$ and a sequence $\{t_n\}$ with $0 < t_n < t_0$ and $t_n \to t_0$. Put $x_n = c(t_n)$, whence $\{x_n\} \subseteq I^-(p)$ and $x_n \to p$. By the lower semicontinuity of distance at (q, p), we have $d(y_n, x_n) > 0$ for some n sufficiently large. Hence $y_n \ll x_n$. On the other hand, by Lemma 4.39 we also have $x_n \ll y_n$. Thus $x_n \ll x_n$, which contradicts Lemma 4.37–(2). Thus $d(q, p) = 0$ and the geodesic λ from q to p is a maximal null segment.

We now show how a timelike curve from $c(0)$ to $c(a)$ of length greater than a may be constructed, contradicting the maximality of the timelike segment c. Consider first the concatenation $\beta = \lambda * c \,|\, [t_0, a]$ which is a future causal curve from q to $c(a)$. Since β consists of a null followed by a timelike geodesic and $L(\beta) = L(c \,|\, [t_0, a]) = a - t_0$, rounding the corner at p produces a causal curve from q to $c(a)$ of length greater than $L(\beta) = a - t_0$. (Take a convex normal neighborhood V centered at p, and join a point close to p on the null geodesic λ to $c(t_0 + \delta)$, for δ sufficiently small, by a timelike geodesic segment lying in V.) Hence, there exists a constant $\epsilon > 0$ such that $d(q, c(a)) \geq (a - t_0) + 4\epsilon$. By Lemma 4.4, we may find a neighborhood $U \subseteq I^+(c(0))$ of q such that

$$d(q', c(a)) \geq a - t_0 + 3\epsilon \tag{4.6}$$

for every q' in U. Choose n sufficiently large that $y_n \in U$ and also that $x_n = c(t_n)$ satisfies $L(c \,|\, [0, t_n]) = d(c(0), c(t_n)) = t_n \geq t_0 - \epsilon$. Fixing this n, let c_1 be a future timelike curve from x_n to y_n guaranteed by Lemma 4.39. Also given y_n, in view of inequality (4.6) we may find a causal curve c_2 from y_n to $c(a)$ with $L(c_2) \geq a - t_0 + 2\epsilon$. Now let $\gamma = c \,|\, [0, t_n] * c_1 * c_2$, which is a future causal curve from $c(0)$ to $c(a)$. On the other hand, we have the length estimate

$$\begin{aligned} L(\gamma) &\geq L(c \,|\, [0, t_n]) + L(c_2) \\ &\geq (t_0 - \epsilon) + (a - t_0 + 2\epsilon) \\ &= a + \epsilon. \end{aligned}$$

But this inequality contradicts the maximality of the timelike segment c. Hence, strong causality must hold at $c(t_0)$. □

CHAPTER 5

EXAMPLES OF SPACE–TIMES

In this chapter we present a variety of examples of space–times. Some of these space–times are important for physical as well as mathematical reasons. In particular, Minkowski space–time, Schwarzschild space–times, Kerr space–times, and Robertson–Walker space–times all have significant physical interpretations.

Minkowski space–time is simultaneously the geometry of special relativity and the geometry induced on each fixed tangent space of an arbitrary Lorentzian manifold. Thus Minkowskian geometry plays the same role for Lorentzian manifolds that Euclidean geometry plays for Riemannian manifolds. Minkowski space–time is sometimes called *flat space–time*. But more generally, any Lorentzian manifold on which the curvature tensor is identically zero is flat.

The Schwarzschild space–times represent the spherically symmetric, empty space–times outside nonrotating, spherically symmetric bodies. Since suns and planets are assumed to be slowly rotating and approximately spherically symmetric, the Schwarzschild space–times may be used to model the gravitational fields outside of these bodies. These space–times may also be used to model the gravitational fields outside of dead (i.e., nonrotating) black holes. The usual coordinates for the Schwarzschild solution outside a massive body are (t, r, θ, ϕ), where t represents a kind of time and r represents a kind of radius [cf. Sachs and Wu (1977a, Chapter 7)]. This metric has a special radius $r = 2m$ associated with it. Points with $r = 2m$ correspond to the surface of a black hole. It was once thought that the metric was singular at $r = 2m$, but it is now known that the usual form of the Schwarzschild metric with $r > 2m$ may be analytically extended to points with $0 < r < 2m$. In fact, there is a maximal analytic extension of Schwarzschild space–time [cf. Kruskal (1960)] which contains an alternative universe lying on the "other side" of the black hole.

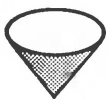
The gravitational fields outside of rotating black holes apparently correspond to the Kerr space–times [cf. Hawking and Ellis (1973, pp. 161, 331), Carter (1971b), O'Neill (1995)]. These space–times represent stationary, axisymmetric metrics outside of rotating objects. The Kerr and Schwarzschild space–times are asymptotically flat and correspond to universes which are empty, apart from one massive body. Thus while these metrics may be reasonable models near a given single massive body, they cannot be used as large scale models for a universe with many massive bodies.

The usual "big bang" cosmological models are based on the Robertson–Walker space–times. These space–times are foliated by a special set of space-like hypersurfaces such that each hypersurface corresponds to an instant of time. The isometry group $I(M)$ of a Robertson–Walker space–time (M, g) acts transitively on these hypersurfaces of constant time. Thus Robertson–Walker universes are spatially homogeneous. Furthermore, they are spatially isotropic in the sense that for each $p \in M$, the subgroup of $I(M)$ fixing p is transitive on the directions at p which are tangential to the hypersurface of constant time through p. In our discussion of Robertson–Walker space–times, we will use Lorentzian warped products $M_0 \times_f H$, as described in Section 3.6. The cosmological assumptions made about Robertson–Walker universes imply that (H, h) is an isotropic Riemannian manifold. Hence the classification of two-point homogeneous Riemannian manifolds yields a classification of all Robertson–Walker space–times. We also show how the results of Section 3.6 may be specialized to construct Lie groups with bi-invariant globally hyperbolic Lorentzian metrics.

5.1 Minkowski Space–time

Minkowski space–time is the manifold $M = \mathbb{R}^n$ together with the metric

$$ds^2 = -dx_1{}^2 + \sum_{i=2}^{n} dx_i{}^2.$$



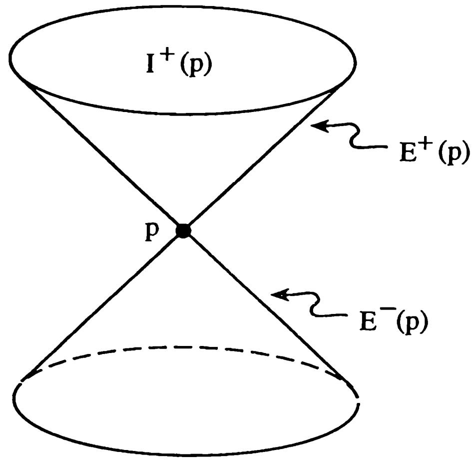


FIGURE 5.1. Let (M, g) be Minkowski space–time. The null cone at p has a future nappe and a past nappe. The future [respectively, past] nappe is also the horismos $E^+(p)$ [respectively, $E^-(p)$] of p. The chronological future $I^+(p)$ is an open convex set bounded by $E^+(p)$. In more general space–times, $I^+(p)$ may fail to be convex but is always open.

This space–time is time oriented by the vector field $\partial/\partial x_1$. It is also globally hyperbolic and hence satisfies all of the causality conditions discussed in Section 3.2.

The geodesics of Minkowski space–time are just the straight lines of the underlying Euclidean space $\mathbb{R}^n$. The affine parametrizations of these geodesics in Minkowski space are even proportional to the usual Euclidean arc length parametrizations in $\mathbb{R}^n$. The null geodesics through a given point p in Minkowski space form an elliptic cone with vertex p. The future directed null geodesics starting at p thus form one nappe of the null cone of p. This nappe forms the boundary in $\mathbb{R}^n$ of an open convex set which is exactly the chronological future $I^+(p)$ of p. In Minkowski space, the causal future $J^+(p)$ of p is the closure of

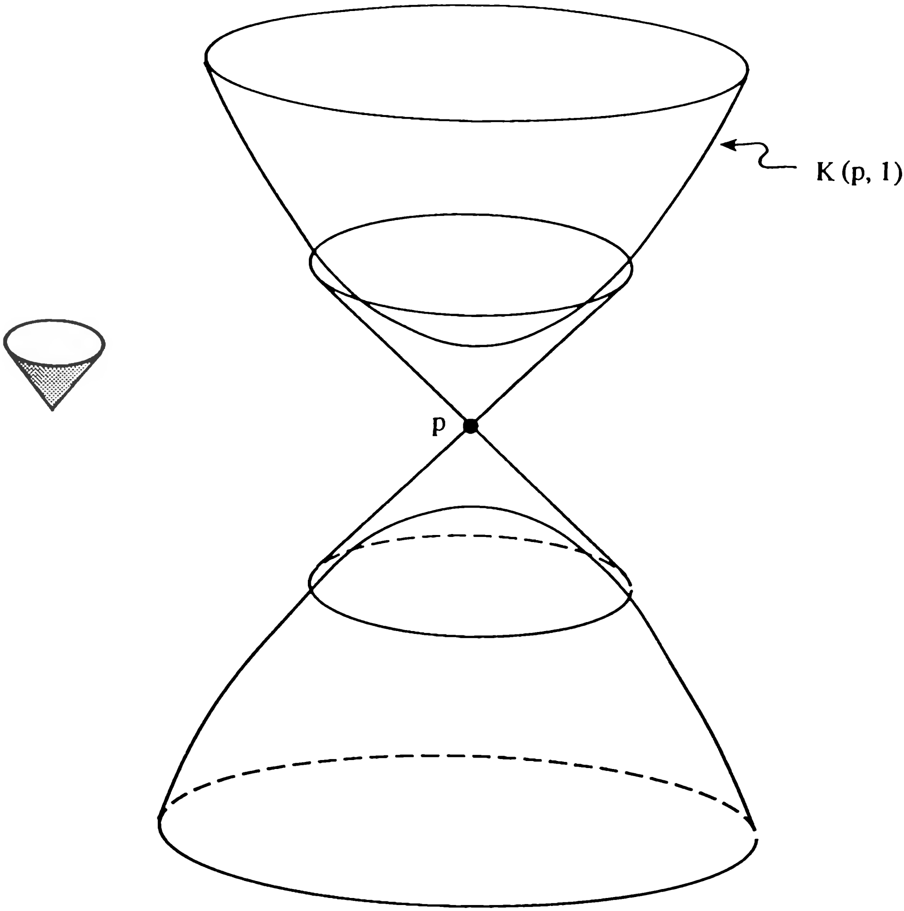


FIGURE 5.2. The unit sphere $K(p, 1)$ corresponding to p is half of a hyperboloid of two sheets. It is *not* compact, and p does *not* lie in the convex open set bounded by $K(p, 1)$.

$I^+(p)$. The future horismos $E^+(p) = J^+(p) - I^+(p)$ is the nappe of the null cone of p corresponding to the future (cf. Figure 5.1).

Minkowski space–time is a Lorentzian product (i.e., a warped product in the sense of Definition 3.51 with $f = 1$). If $\mathbb{R}$ is given the negative definite

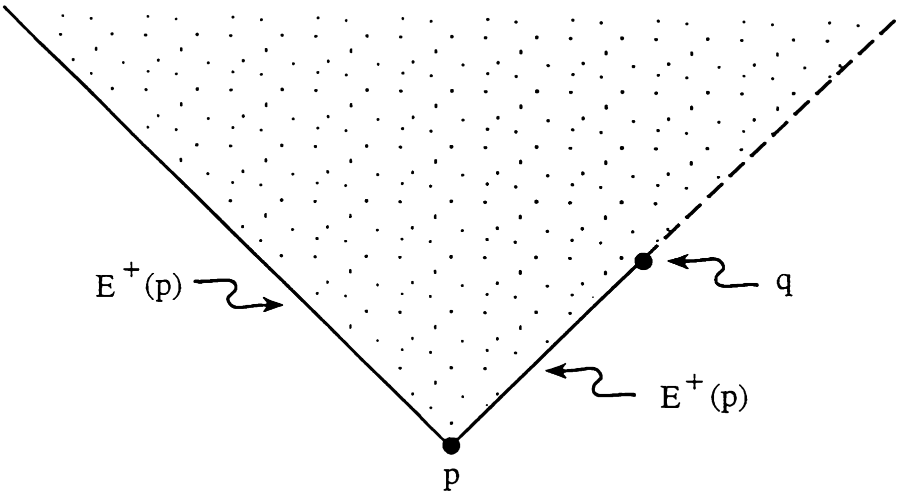


FIGURE 5.3. Two-dimensional Minkowski space–time with one point q removed is shown. The future horismos $E^+(p)$ of p is an "L" shaped figure consisting of a half-closed line and a half-open line segment. The causal future $J^+(p)$ is the union of $I^+(p)$ and $E^+(p)$. Notice that $J^+(p)$ is *not* a closed set nor is $J^+(p)$ equal to the closure of $I^+(p)$.

metric $-dt^2$ and $\mathbb{R}^{n-1}$ is given the usual Euclidean metric g_0, then $(\mathbb{R}^n = \mathbb{R} \times \mathbb{R}^{n-1}, -dt^2 \oplus g_0)$ is the n-dimensional Minkowski space–time.

Consider two points $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ in Minkowski space–time. The chronological relation $p \ll q$ holds whenever $p_1 < q_1$ and $(p_1 - q_1)^2 > (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2$ in $\mathbb{R}$. If $p \ll q$, then the distance from p to q is given by

$$d(p, q) = \left[(p_1 - q_1)^2 - \sum_{i=2}^{n} (p_i - q_i)^2 \right]^{\frac{1}{2}}$$

The "unit sphere" in Minkowski space–time centered at p is then $K(p, 1) = \{q \in M : d(p, q) = 1\}$. However, this set is actually one sheet of a hyperboloid of two sheets (cf. Figure 5.2).

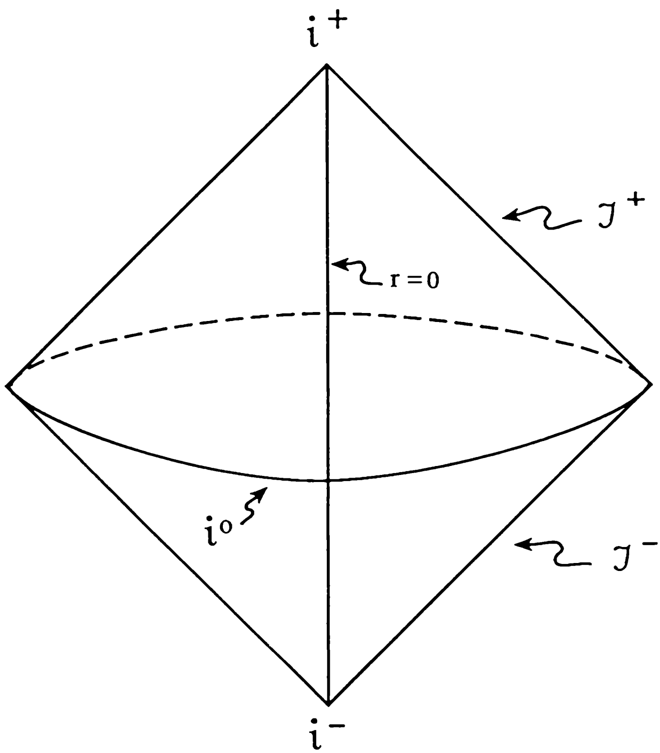


FIGURE 5.4. Minkowski space–time is conformal to the open set enclosed by the two null cones indicated. The vertices i^+ and i^- correspond to timelike infinity. All future directed timelike geodesics go from i^- to i^+. The sets $\mathfrak{I}^+$ and $\mathfrak{I}^-$ represent future and past null infinity. Topologically, $\mathfrak{I}^+$ and $\mathfrak{I}^-$ are each $\mathbb{R} \times S^{n-2}$. The intersection of the two null cones is a set which is identified to a single point i^0. The point i^0 is called *spacelike infinity*.

If we remove a point from Minkowski space–time, then it is no longer causally simple and hence no longer globally hyperbolic (cf. Figure 5.3).

It is possible to conformally map all of Minkowski space–time onto a small open set about the origin. This is illustrated in Figure 5.4 [cf. Penrose (1968, p. 178), Hawking and Ellis (1973, p. 123)].

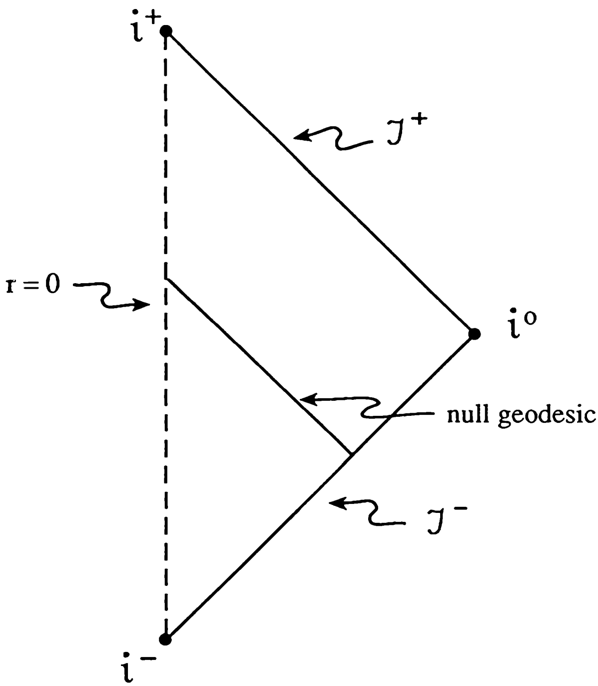


FIGURE 5.5. The Penrose diagram for Minkowski space–time is shown.

Minkowski space–time and many other important space–times may be represented by *Penrosė diagrams*. A Penrose diagram is a two-dimensional representation of a spherically symmetric space–time. The radial null geodesics are represented by null geodesics at $\pm 45°$. Dotted lines represent the origin ($r = 0$) of polar coordinates. Points corresponding to smooth boundary points (cf. Section 12.5) which are not singularities are represented by single lines. Double lines represent irremovable singularities (Figure 5.5; cf. Figure 4.2 of a Reissner–Nordström space–time with $e^2 = m^2$ for an example).

5.2 Schwarzschild and Kerr Space–times

In this section we describe the four-dimensional Schwarzschild and Kerr solutions to the Einstein equations. Let $\mathbb{R}^4$ be given coordinates (t, r, θ, ϕ), where (r, θ, ϕ) are the usual spherical coordinates on $\mathbb{R}^3$. Given a positive constant m, the exterior Schwarzschild space–time is defined on the subset $r > 2m$ of $\mathbb{R}^4$, a subset which is topologically $\mathbb{R}^2 \times S^2$. The Schwarzschild metric for the region $r > 2m$ is given in (t, r, θ, ϕ) coordinates by the formula

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dt^2 + \left(1 - \frac{2m}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2).$$

Each element of the rotation group $SO(3)$ for $\mathbb{R}^3$ induces a motion of the Schwarzschild solution. Given $\psi \in SO(3)$, a motion $\overline{\psi}$ of Schwarzschild space–time may be defined by setting $\overline{\psi}(t, r, \theta, \phi) = (t, \psi(r, \theta, \phi))$. Thus at a fixed instant t in time, the exterior Schwarzschild space–time is spherically symmetric. The metric for this space–time is also invariant under the time translation $t \to t + a$. The coordinate vector field $\partial/\partial t$ is a timelike Killing vector field which is a gradient, and the metric is said to be *static*. This space–time is also Ricci flat (i.e., $\text{Ric} = 0$). Using the Einstein equations (cf. Chapter 2), it follows that the energy-momentum tensor for the exterior Schwarzschild space–time vanishes. Thus this space–time is empty.

The exterior Schwarzschild space–time may be regarded as a Lorentzian warped product (cf. Section 3.6). For let $M = \{(t, r) \in \mathbb{R}^2 : r > 2m\}$ be given the Lorentzian metric

$$g = -\left(1 - \frac{2m}{r}\right) dt^2 + \left(1 - \frac{2m}{r}\right)^{-1} dr^2,$$

and let $H = S^2$ be given the usual Riemannian metric h of constant sectional curvature one induced by the inclusion $S^2 \to \mathbb{R}^3$. Define $f : M \to \mathbb{R}$ by $f(t, r) = r^2$. Then $(M \times_f H, \overline{g})$ is the exterior Schwarzschild space–time, where $\overline{g} = g \oplus fh$.

Physically, the exterior Schwarzschild solution represents the gravitational field outside of a nonrotating spherically symmetric massive object. Comparison with the Newtonian theory [cf. Einstein (1916, p. 819), Pathria (1974, p. 217)] shows that m can be identified with the gravitational mass of the massive body. The solution is not valid in the interior of the body.

The above form of the exterior Schwarzschild metric appears to have a singularity at $r = 2m$. However, this is not a true singularity; the exterior

Schwarzschild solution may be analytically continued across the surface given by the equation $r = 2m$.

Kruskal (1960) investigated the maximal analytic extension of Schwarzschild space–time. Suppressing θ and ϕ, the following two-dimensional representation of this maximal extension may be given (cf. Figure 5.6).

The gravitational field outside of a rotating black hole will not correspond to the Schwarzschild solution. The generally accepted solutions of the Einstein equations for rotating black holes are Kerr solutions. In Boyer and Lindquist coordinates (t, r, θ, ϕ) the Kerr metrics are given by [cf. Hawking and Ellis (1973, p. 161), O'Neill (1995)]

$$ds^2 = \rho^2 \left(\frac{dr^2}{\Delta} + d\theta^2 \right) + (r^2 + a^2) \sin^2 \theta d\phi^2 - dt^2 + \frac{2mr}{\rho^2} (a \sin^2 \theta d\phi - dt)^2$$

where $\rho^2 = r^2 + a^2 \cos^2 \theta$ and $\Delta = r^2 - 2mr + a^2$. The constant m represents the mass, and the constant ma represents the angular momentum of the black hole [cf. Boyer and Price (1965), Boyer and Lindquist (1967)]. Tomimatsu and Sato (1973) have given a series of exact solutions which include the Kerr solutions as special cases.

5.3 Spaces of Constant Curvature

It is known that two Lorentzian manifolds of the same dimension which have constant sectional curvature K are locally isometric [cf. Wolf (1974, p. 69)]. Thus any Lorentzian manifold of constant sectional curvature zero is locally isometric to Minkowski space–time. In this section we will consider Lorentzian model spaces which have constant nonzero sectional curvature.

We first define $\mathbb{R}^n_s$ to be the standard semi-Euclidean space of signature $(-, \ldots, -, +, \ldots, +)$, where there are s negative eigenvalues and $n - s$ positive eigenvalues. Hence the semi-Euclidean metric on $\mathbb{R}^n_s$ is given by

$$ds^2 = -\sum_{i=1}^{s} dx_i{}^2 + \sum_{i=s+1}^{n} dx_i{}^2.$$

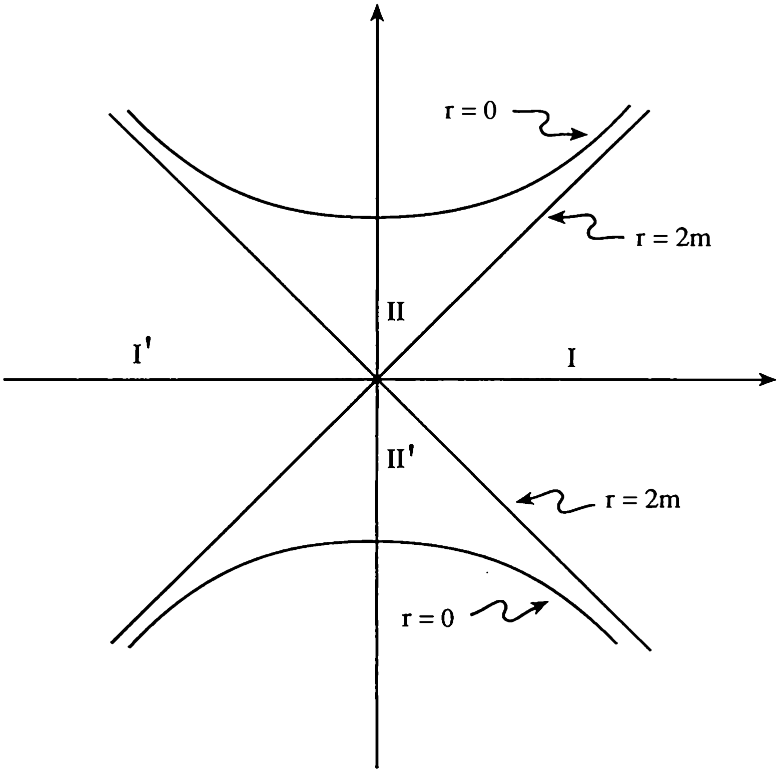


FIGURE 5.6. The Kruskal diagram for the maximal analytic extension of the exterior Schwarzschild space–time is shown. The extended space–time is the connected nonconvex region $\mathrm{I} \cup \mathrm{II} \cup \mathrm{I}' \cup \mathrm{II}'$ bounded by the hyperbola corresponding to $r = 0$. The points of this hyperbola are the true singularities of this space–time. The lines at $\pm 45°$ separate the space–time into four regions. Region I corresponds to the exterior Schwarzschild solution. Region II is the "interior" of a nonrotating black hole. Region I′ is isometric to region I and corresponds to an alternative universe on the "other side" of the black hole. There is no nonspacelike curve from region I to region I′.

In particular, $\mathbb{R}^n_1$ is n-dimensional Minkowski space–time. We also define for $r > 0$ [cf. Wolf (1974, Section 5.2)]

$$S^n_1 = \{x \in \mathbb{R}^{n+1}_1 : -{x_1}^2 + {x_2}^2 + \cdots + x^2_{n+1} = r^2\}$$

and

$$H^n_1 = \{x \in \mathbb{R}^{n+1}_2 : -{x_1}^2 - {x_2}^2 + {x_3}^2 + \cdots + x^2_{n+1} = -r^2\}.$$

Topologically, S^n_1 is $\mathbb{R}^1 \times S^{n-1}$ and H^n_1 is $S^1 \times \mathbb{R}^{n-1}$ [cf. Wolf (1974, p. 68)]. The semi-Euclidean metric on $\mathbb{R}^{n+1}_1$ (respectively, $\mathbb{R}^{n+1}_2$) induces a Lorentzian metric of constant sectional curvature $K = r^{-2}$ (respectively, $K = -r^{-2}$) on S^n_1 (respectively, H^n_1). The space–time S^n_1 is a Lorentzian analogue of the usual Riemannian spherical space of radius r and has positive curvature r^{-2}. The universal covering manifold $\widetilde{H}^n_1$ of H^n_1 is topologically $\mathbb{R}^n$ and is thus a Lorentzian analogue of the usual Riemannian hyperbolic space of negative curvature $-r^{-2}$.

Definition 5.1. (*de Sitter and anti–de Sitter Space–times*) Let S^n_1 and H^n_1 be defined as above. Then S^n_1 is called *de Sitter space–time*, and the universal covering $\widetilde{H}^n_1$ of H^n_1 is called (*universal*) *anti–de Sitter space–time.*

Remark 5.2.

(1) S^n_1 is simply connected for $n > 2$ and $\pi_1(S^2_1) = \mathbb{Z}$.

(2) S^n_1 is globally hyperbolic and geodesically complete.

(3) H^n_1 is nonchronological since $\gamma(t) = (r\cos t, r\sin t, 0, \ldots, 0)$ is a closed timelike curve. Also $\widetilde{H}^n_1$, while strongly causal, is not globally hyperbolic.

The de Sitter space–time represented in Figure 5.7 may be covered by global coordinates (t, χ, θ, ϕ) with $-\infty < t < \infty$, $0 \leq \chi \leq \pi$, $0 \leq \theta \leq \pi$, and $0 \leq \phi \leq 2\pi$. Here t is the coordinate on $\mathbb{R}$ and (χ, θ, ϕ) represent coordinates on S^3 [cf. Hawking and Ellis (1973, pp. 125, 136)]. In these coordinates, the metric for de Sitter space–time of constant positive sectional curvature $1/r^2$ is given by

$$ds^2 = -dt^2 + r^2 \cosh^2\left(\frac{t}{r}\right)\left[d\chi^2 + \sin^2\chi\,(d\theta^2 + \sin^2\theta\, d\phi^2)\right].$$

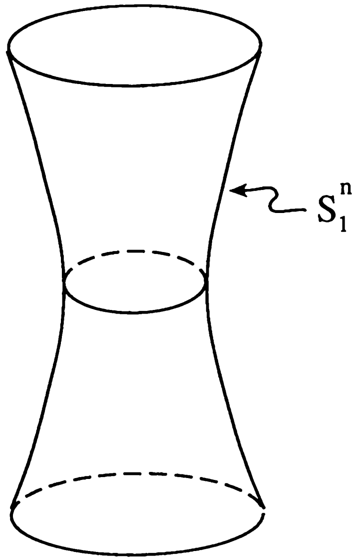


FIGURE 5.7. The n-dimensional de Sitter space–time with positive constant sectional curvature r^{-2} is the set $-{x_1}^2+{x_2}^2+\cdots+x_{n+1}^2 = r^2$ in Minkowski space–time $\mathbb{R}_1^{n+1}$. The geodesics of S_1^n lie on the intersection of S_1^n with the planes through the origin of $\mathbb{R}_1^{n+1}$.

This may be reinterpreted as a Lorentzian warped product metric (cf. Section 3.6) as follows. Let $f : \mathbb{R} \to (0, \infty)$ be given by $f(t) = r^2 \cosh^2(t/r)$, and let S^3 be given the usual complete Riemannian metric of constant sectional curvature one. Then the de Sitter space–time described in local coordinates as above is the warped product $(\mathbb{R} \times S^3, -dt^2 \oplus fh)$.

Universal anti–de Sitter space–time of curvature $K = -1$ may be given coordinates (t', r, θ, ϕ) for which the metric has the form

$$ds^2 = -\cosh^2(r)(dt')^2 + dr^2 + \sinh^2(r)(d\theta^2 + \sin^2\theta\, d\phi^2)$$

[cf. Hawking and Ellis (1973, pp. 131, 136)]. Regarding $-(dt')^2$ as a negative definite metric on $\mathbb{R}$ and $dr^2 + \sinh^2(r)(d\theta^2 + \sin^2\theta\, d\phi^2)$ as the complete

Riemannian metric h of constant negative sectional curvature -1 on the hyperbolic three-space $H = \mathbb{R}^3$, this space–time may be represented as a warped product of the form $(\mathbb{R} \times_f H, -f dt^2 \oplus h)$, where the warping function is defined on the Riemannian factor H (cf. Remark 3.53).

5.4 Robertson–Walker Space–times

In this section we discuss Robertson–Walker space–times in the framework of Lorentzian warped products. These space–times include the Einstein static universe and the big bang cosmological models of general relativity. In order to give a precise definition of a Robertson–Walker space–time, it is necessary to first recall some concepts from the theory of two-point homogeneous Riemannian manifolds and isotropic Riemannian manifolds.

Let (H, h) be a Riemannian manifold. Denote by $I(H)$ the isometry group of (H, h) and by $d_0 : H \times H \to \mathbb{R}$ the Riemannian distance function of (H, h).

Definition 5.3. (*Homogeneous and Two-Point Homogeneous Manifolds*) The Riemannian manifold (H, h) is said to be *homogeneous* if $I(H)$ acts transitively on H, i.e., given any $p, q \in H$, there is an isometry $\phi \in I(H)$ with $\phi(p) = q$. Further, (H, h) is said to be *two-point homogeneous* if given any $p_1, q_1, p_2, q_2 \in H$ with $d_0(p_1, q_1) = d_0(p_2, q_2)$, there is an isometry $\phi \in I(H)$ with $\phi(p_1) = p_2$ and $\phi(q_1) = q_2$.

Since it is possible to choose $p_i = q_i$ for $i = 1, 2$, a two-point homogeneous Riemannian manifold is also homogeneous. Two-point homogeneous spaces were first studied by Busemann (1942) in the more general setting of locally compact metric spaces. Wang (1951, 1952) and Tits (1955) classified two-point homogeneous Riemannian manifolds.

Notice that in Definition 5.3 it is not required that (H, h) be a complete Riemannian manifold. Nonetheless, homogeneous Riemannian manifolds have the important basic property of always being complete.

Lemma 5.4. *If (H, h) is a homogeneous Riemannian manifold, then (H, h) is complete.*

Proof. By the Hopf–Rinow Theorem, it suffices to show that (H, h) is geodesically complete. Thus suppose that $c : [a, 1) \to H$ is a unit speed geodesic which is not extendible to $t = 1$. Choosing any $p \in H$, we may find a constant $\alpha > 0$ such that any unit speed geodesic starting at p has length $l \geq \alpha$. Set $\delta = \min\{\alpha/2, (1 - a)/2\} > 0$. Since isometries preserve geodesics, it follows from the homogeneity of (H, h) that any unit speed geodesic starting at $c(1 - \delta)$ may be extended to a geodesic of length $l \geq 2\delta$. In particular, c may be extended to a geodesic $c : [a, 1 + \delta) \to H$, in contradiction to the inextendibility of c to $t = 1$. □

Remark 5.5. It is important to note that the conclusion of Lemma 5.4 is false in general for homogeneous Lorentzian manifolds [cf. Wolf (1974, p. 95), Marsden (1973)].

We now recall the concept of an isotropic Riemannian manifold. Given $p \in (H, h)$, the *isotropy group* $I_p(H)$ of (H, h) at p is the closed subgroup $I_p(H) = \{\phi \in I(H) : \phi(p) = p\}$ of $I(H)$ consisting of all isometries of (H, h) which fix p. Given any $\phi \in I_p(H)$, the differential ϕ_{*_p} maps T_pH onto T_pH since $\phi(p) = p$. As $h(\phi_* v, \phi_* v) = h(v, v)$ for any $v \in T_pH$, the differential ϕ_{*_p} also maps the unit sphere $S_pH = \{v \in T_pM : h(v, v) = 1\}$ in T_pH onto itself.

Definition 5.6. (*Isotropic Riemannian Manifold*) A Riemannian manifold (H, h) is said to be *isotropic at p* if $I_p(H)$ acts transitively on the unit sphere S_pH of T_pH, i.e., given any $v, w \in S_pH$, there is an isometry $\phi \in I_p(H)$ with $\phi_* v = w$. The Riemannian manifold (H, h) is said to be *isotropic* if it is isotropic at every point.

We now show that the class of isotropic Riemannian manifolds coincides with the class of two-point homogeneous Riemannian manifolds [cf. Wolf (1974, p. 289)].

Proposition 5.7. *A Riemannian manifold (H, h) is isotropic iff it is two-point homogeneous.*

Proof. Recall that d_0 denotes the Riemannian distance function of (H, h). First suppose that (H, h) is isotropic. Then for each $p \in H$ and each inextendible geodesic $c : (a, b) \to H$ with $c(0) = p$, there is an isometry $\phi \in I_p(H)$

with $\phi_* c'(0) = -c'(0)$. Hence by geodesic uniqueness, $\phi(c(t)) = c(-t)$ for all $t \in (a, b)$. This implies that the length of $c|\,(a, 0]$ equals the length of $c|\,[0, b)$. Since p may be taken to be any point of the geodesic c, it follows that $a = -\infty$ and $b = +\infty$. Thus (H, h) is geodesically complete. Hence by the Hopf–Rinow Theorem, given any two points $p_1, p_2 \in H$, there is a geodesic segment c_0 of minimal length $d_0(p_1, p_2)$ from p_1 to p_2. Let p be the midpoint of c_0. As (H, h) is isotropic, there is an isometry $\phi \in I_p(H)$ which reverses c_0. It follows that $\phi(p_1) = p_2$. Hence (H, h) is homogeneous. It remains to show that if $p_1, q_1, p_2, q_2 \in H$ with $d_0(p_1, q_1) = d_0(p_2, q_2) > 0$ are given, we may find an isometry $\phi \in I(H)$ with $\phi(p_1) = p_2$ and $\phi(q_1) = q_2$. Choose minimal unit speed geodesics c_1 from p_1 to q_1 and c_2 from p_2 to q_2. Since (H, h) is homogeneous, we may first find an isometry $\psi \in I(H)$ with $\psi(p_1) = p_2$. Then as (H, h) is isotropic, we may find $\eta \in I_{p_2}(H)$ with $\eta_*((\psi \circ c_1)'(0)) = c_2'(0)$. It follows that $\phi = \eta \circ \psi$ is the required isometry.

Now suppose that (H, h) is two-point homogeneous. Fix any $p \in M$, and let U be a convex normal neighborhood based at p. Choose $\alpha > 0$ such that $\exp_p(v) \in U$ for all $v \in T_pH$ with $h(v, v) \leq \alpha$. Now let $v, w \in T_pH$ be any pair of nonzero tangent vectors with $h(v, v) = h(w, w) < \alpha/2$. Set $q_1 = \exp_p v$ and $q_2 = \exp_p w$. Then $q_1, q_2 \in U$ and $d(p, q_1) = \sqrt{h(v, v)} = \sqrt{h(w, w)} = d(p, q_2)$. Since (H, h) is two-point homogeneous, there is thus an isometry $\phi \in I(H)$ with $\phi(p) = p$ and $\phi(q_1) = q_2$. It follows that $\phi_* v = w$. The linearity of $\eta_{*_p} : T_pH \to T_pH$ for any $\eta \in I_p(H)$ then implies that $I_p(H)$ acts transitively on S_pH. Thus (H, h) is isotropic at p. As the same argument clearly holds for all $p \in H$, it follows that (H, h) is isotropic as required. □

Corollary 5.8. *Any isotropic Riemannian manifold is homogeneous and complete.*

Remark 5.9. (1) The two-point homogeneous Riemannian manifolds are well known [cf. Wolf (1974, pp. 290–296)]. In particular, the odd-dimensional two-point homogeneous (hence isotropic) Riemannian manifolds are just the odd-dimensional Euclidean, hyperbolic, spherical, and elliptic spaces [cf. Wang (1951, p. 473)].

(2) Astronomical observations indicate that the spatial universe is approxi-

mately spherically symmetric about the earth. This suggests that the spatial universe should be modeled as a three-dimensional isotropic Riemannian manifold. Hence the possibilities are limited to the Euclidean, hyperbolic, spherical, and elliptic spaces. However, if one only assumes local isotropy, there are more possibilities [cf. Misner, Thorne, and Wheeler (1973, pp. 713–725)].
(3) Any three-dimensional isotropic Riemannian manifold (H, h) has constant sectional curvature, and also $\dim I(H) = 6$ [cf. Walker (1944)].

We are now ready to define Robertson–Walker space–times using Lorentzian warped products and isotropic Riemannian manifolds.

Definition 5.10. (*Robertson–Walker Space–time*) A *Robertson–Walker space–time* (M, g) is any Lorentzian manifold which can be written in the form of a Lorentzian warped product $(M_0 \times_f H, g)$ with $M_0 = (a, b)$, $-\infty \leq a < b \leq +\infty$, given the negative definite metric $-dt^2$, with (H, h) an isotropic Riemannian manifold, and with warping function $f : M_0 \to (0, \infty)$.

In the notation of Section 3.6, we thus have $g = -dt^2 \oplus fh$ and $M_0 \times_f H$ is also topologically the product $M_0 \times H$. Letting $d\sigma^2$ denote the Riemannian metric h for H and defining $S(t) = \sqrt{f(t)}$, the Lorentzian metric g for $M_0 \times_f H$ may be rewritten in the more familiar form

$$ds^2 = -dt^2 + S^2(t)d\sigma^2.$$

The map $\pi : M_0 \times_f H \to \mathbb{R}$ given by $\pi(t, x) = t$ is a smooth time function on $M_0 \times_f H$ so that the Lorentzian manifold $M_0 \times_f H$, of Definition 5.10 actually is a (stably causal) space–time. Also each level surface $\pi^{-1}(c)$ of the map $\pi : M_0 \times_f H \to M_0 \subseteq \mathbb{R}$ is an isotropic Riemannian manifold which is homothetic to (H, h). Furthermore, the isometry group $I(H)$ of (H, h) may be identified with a subgroup $\widetilde{I}(H)$ of $I(M_0 \times_f H)$ as follows. Given $\phi \in I(H)$, define $\widetilde{\phi} \in \widetilde{I}(H)$ by $\widetilde{\phi}(r, h) = (r, \phi(h))$ for all $(r, h) \in M_0 \times H$. With this definition, $\widetilde{I}(H)$ restricted to the level surfaces $\pi^{-1}(c)$ of π acts transitively on each level surface.

Since all isotropic Riemannian manifolds are complete, Theorem 3.66 implies that all Robertson–Walker space–times are globally hyperbolic. From

Theorem 3.69 we also know that every level surface $\pi^{-1}(c) = \{c\} \times H$ is a Cauchy surface for $M_0 \times_f H$.

Next to Minkowski space $\mathbb{R}^n = \mathbb{R} \times \mathbb{R}^{n-1}$ itself, the Einstein static universe is the simplest example of a Robertson–Walker space–time.

Example 5.11. (*Einstein Static Universe*) *Let $M_0 = \mathbb{R}$ with the negative definite metric $-dt^2$, and let $H = S^{n-1}$ with the standard spherical Riemannian metric. If $f : \mathbb{R} \to (0, \infty)$ is the trivial warping function $f = 1$, then the product Lorentzian manifold $M = M_0 \times H = M_0 \times_f H$ is the n-dimensional Einstein static universe. If $n = 2$, then M is the cylinder $\mathbb{R} \times S^1$ with flat metric $-dt^2 + d\theta^2$. If $n \geq 3$, then this metric for $M = \mathbb{R} \times S^{n-1}$ is not flat since S^{n-1} has constant positive sectional curvature $K = 1$.*

For the rest of this section we restrict our attention to four-dimensional Robertson–Walker space–times. By Remark 5.9, these are warped products $M_0 \times_f H$, where (H, h) is Euclidean, hyperbolic, spherical, or elliptic of dimension three. In the first two cases, H is topologically $\mathbb{R}^3$. In the third case $H = S^3$, and in the last case H is the real projective three-space $\mathbb{R}P^3$. We thus have the following

Corollary 5.12. *All four-dimensional Robertson–Walker space–times are topologically either $\mathbb{R}^4$, $\mathbb{R} \times S^3$, or $\mathbb{R} \times \mathbb{R}P^3$.*

Also by Remark 5.9, the sectional curvature K of (H, h) is constant. If K is nonzero, the metric may be rescaled to be of the form $ds^2 = -dt^2 + S^2(t)d\sigma^2$ on M so that K is either identically $+1$ or -1. This is the form of the metric usually studied in general relativity.

In physics, cosmological models are built from four-dimensional Robertson–Walker space–times assumed to be filled with a perfect fluid. The Einstein equations (cf. Chapter 2) are then used to find the form of the above warping function $S^2(t)$. Among the models this technique yields are the big bang cosmological models [cf. Hawking and Ellis (1973, pp. 134–138)]. These models depend on the energy density μ and pressure p of the perfect fluid as well as the value of the cosmological constant Λ in the Einstein equations. In the big bang cosmological models, the inextendible nonspacelike geodesics are all past incomplete. The stability of this incompleteness under metric perturbations will be considered in Section 7.3. Astronomical observations of clusters of galaxies indicate that distant clusters of galaxies are receding from us. This expansion of the universe suggests the existence of a "big bang" in the past and also suggests that the universe is a warped product with a nontrivial warping function rather than simply a Lorentzian product. Observations of blackbody radiation support these ideas [cf. Hawking and Ellis (1973, Chapter 10)].

5.5 Bi-Invariant Lorentzian Metrics on Lie Groups

The purpose of this section is to show how Theorems 3.67 and 3.68 of Section 3.6 may be used to construct a large class of Lie groups admitting globally hyperbolic, bi-invariant Lorentzian metrics.

We first summarize some basic facts from the elementary theory of Lie groups. Details may be found in a lucid exposition by Milnor (1963, Part IV) or at a more advanced level in Helgason (1978, Chapter 2). A Lie group is a group G which is also an analytic manifold such that the mapping $(g, h) \to gh^{-1}$ from $G \times G \to G$ is analytic. This multiplication induces left and right translation maps L_g, R_g for each $g \in G$, given respectively by $L_g(h) = gh$ and $R_g(h) = hg$. A Riemannian or Lorentzian metric $\langle\ ,\ \rangle$ for G is then said to be *left invariant* (respectively, *right invariant*) if $\langle L_{g*}v, L_{g*}w\rangle = \langle v, w\rangle$ (respectively, $\langle R_{g*}v, R_{g*}w\rangle = \langle v, w\rangle$) for all $g \in G$ and $v, w \in TG$. A metric which is both left and right invariant is said to be *bi-invariant*. By an averaging procedure involving the Haar integral, any compact Lie group may be given a bi-invariant metric [cf. Milnor (1963, p. 112)]. In fact, the Haar integral may be used to produce a bi-invariant Riemannian metric for G from any left invariant Riemannian metric for G. Any Lie group may be equipped with a left invariant Riemannian (or Lorentzian) metric by starting with a positive definite inner product (respectively, inner product of signature $n-2$) $\langle\ ,\ \rangle|_e$ on the tangent space T_eG to G at the identity element $e \in G$, then defining $\langle\ ,\ \rangle|_g : T_gG \times T_gG \to \mathbb{R}$ by

$$\langle v, w\rangle\big|_g = \langle L_{g^{-1}*}v, L_{g^{-1}*}w\rangle\big|_e . \tag{5.1}$$

Thus any compact Lie group is furnished with a large supply of bi-invariant Riemannian metrics.

On the other hand, while (5.1) equips any Lie group with left-invariant Lorentzian metrics, the standard Haar integral averaging procedure used for Riemannian metrics fails to preserve signature $(-,+,\cdots,+)$, so it cannot be used to convert left-invariant Lorentzian metrics into bi-invariant Lorentzian metrics.

But we will see shortly that a large class of bi-invariant Lorentzian metrics may be constructed for noncompact Lie groups of the form $\mathbb{R} \times G$, where G is any Lie group admitting a bi-invariant Riemannian metric.

Before giving the construction, we need to discuss product Lie groups briefly. Let G and H be two Lie groups. The product manifold $G \times H$ is then turned into a Lie group by defining the multiplication by

$$(g_1, h_1) \times (g_2, h_2) = (g_1 g_2, h_1 h_2). \tag{5.2}$$

It is immediate from (5.2) that if $\sigma = (g, h) \in G \times H$, then the translation maps $L_\sigma,\ R_\sigma : G \times H \to G \times H$ are given by $L_\sigma = (L_g, L_h)$, and $R_\sigma = (R_g, R_h)$, i.e., $L_\sigma(g_1, h_1) = (L_g g_1, L_h h_1)$, etc. Recall that $T_\sigma(G \times H) \cong T_g G \times T_h H$. It is straighforward to check that for any $\sigma \in G \times H$ and any tangent vector $\xi = (v, w) \in T_\sigma(G \times H) \cong T_g G \times T_h H$, one has

$$L_{\sigma_*}\xi = (L_{g_*}v, L_{h_*}w) \tag{5.3}$$

and

$$R_{\sigma_*}\xi = (R_{g_*}v, R_{h_*}w)\,. \tag{5.4}$$

Now if $\langle\ ,\ \rangle_1$ is a Lorentzian metric for G and $\langle\ ,\ \rangle_2$ is a Riemannian metric for H, the product metric $\langle\langle\ ,\ \rangle\rangle = \langle\ ,\ \rangle_1 \oplus \langle\ ,\ \rangle_2$ is a Lorentzian metric for $G \times H$. Explicitly, recalling Definition 3.51, we have for tangent vectors $\xi_1 = (v_1, w_1)$ and $\xi_2 = (v_2, w_2)$ in $T_\sigma(G \times H)$ the formula

$$\langle\langle \xi_1, \xi_2 \rangle\rangle = \langle v_1, v_2 \rangle_1 + \langle w_1, w_2 \rangle_2.$$

It is then immediate from (5.3) and (5.4) that if $\langle\ ,\ \rangle_1$ is a bi-invariant Lorentzian metric for G and $\langle\ ,\ \rangle_2$ is a bi-invariant Riemannian metric for H, then $\langle\langle\ ,\ \rangle\rangle$ is a bi-invariant Lorentzian metric for $G \times H$. To summarize,

Proposition 5.13. *Let $(G, \langle\ ,\ \rangle_1)$ be a Lie group equipped with a bi-invariant Lorentzian metric, and let $(H, \langle\ ,\ \rangle_2)$ be a Lie group equipped with a bi-invariant Riemannian metric. Then the product metric $\langle\langle\ ,\ \rangle\rangle = \langle\ ,\ \rangle_1 \oplus \langle\ ,\ \rangle_2$ is a bi-invariant Lorentzian metric for the product Lie group $G \times H$. Hence $(G \times H, \langle\langle\ ,\ \rangle\rangle)$ is a Lorentzian symmetric space and, in particular, is geodesically complete.*

Proof. It is only necessary to prove the last statement which is a standard fact in Lie group theory. Recall that we must show that for each $\sigma \in G \times H$, there exists an isometry $I_\sigma : G \times H \to G \times H$ which fixes σ and reverses the geodesics through σ. That is, if γ is a geodesic in $G \times H$ with $\gamma(0) = \sigma$, we must show that $I_\sigma(\gamma(t)) = \gamma(-t)$ for all t. This is equivalent to showing that $I_{\sigma_*} : T_\sigma(G \times H) \to T_\sigma(G \times H)$ is the map $I_{\sigma_*}(\xi) = -\xi$ and also implies that ${I_\sigma}^2 = \text{Id}$.

We will follow the proof given in Milnor (1963, pp. 109, 112). First, if we denote the identity element of $G \times H$ by e and define a map $I_e : G \times H \to G \times H$ by $I_e(\sigma) = \sigma^{-1}$, then $I_{e_*} : T_e(G \times H) \to T_e(G \times H)$ is given by $I_{e_*}(v) = -v$. Thus $I_{e_*} : T_e(G \times H) \to T_e(G \times H)$ is an isometry of $T_e(G \times H)$. To see that I_{e_*} is an isometry of any other tangent space $T_\sigma(G \times H) \to T_{\sigma^{-1}}(G \times H)$ and hence that $I_e : G \times H \to G \times H$ is an isometry, we simply note that

$$I_e = R_{\sigma^{-1}} I_e L_{\sigma^{-1}}.$$

Since $\langle\langle\ ,\ \rangle\rangle$ is bi-invariant, all left and right translation maps are isometries. Then as

$$I_{e_*}\big|_\sigma = R_{\sigma_*^{-1}}\big|_e I_{e_*}\big|_e L_{\sigma_*^{-1}}\big|_\sigma$$

and $I_{e_*}|_e$ is an isometry of $T_e(G \times H)$, it follows that $I_{e_*} : T_\sigma(G \times H) \to T_{\sigma^{-1}}(G \times H)$ is also an isometry. The map $I_e : G \times H \to G \times H$ is thus the required geodesic symmetry at e.

We define the geodesic symmetry I_σ for any $\sigma \in G$ by setting $I_\sigma = R_\sigma I_e R_{\sigma^{-1}}$. Since R_σ and $R_{\sigma^{-1}}$ are isometries by the bi-invariance of $\langle\langle\ ,\ \rangle\rangle$ and we have just shown that I_e is an isometry, it follows that $I_\sigma : G \times H \to G \times H$ is an isometry, and obviously $I_\sigma(\sigma) = \sigma$ since $I_\sigma(h) = \sigma h^{-1} \sigma$. Finally, for any $\xi \in T_\sigma(G \times H)$ we have

$$\begin{aligned} I_{\sigma_*}\xi &= R_{\sigma_*}\left(I_{e_*}\left(R_{\sigma_*^{-1}}\xi\right)\right) \\ &= R_{\sigma_*}\left(-R_{\sigma_*^{-1}}\xi\right) \qquad \text{since } R_{\sigma_*^{-1}}\xi \in T_e(G \times H) \\ &= -R_{\sigma_*}R_{\sigma_*^{-1}}\xi \\ &= -\left(R_\sigma R_{\sigma^{-1}}\right)_*\xi = -\xi. \end{aligned}$$

Thus I_σ reverses geodesics at σ as required. We have therefore shown that $G \times H$ is a symmetric space.

It may be shown that any symmetric space is geodesically complete as follows. Let γ be a geodesic in M, and set $p = \gamma(0)$. Supposing that $q = \gamma(A)$ is defined, one may derive the formula [cf. Milnor (1963, p. 109)]

$$I_q I_p(\gamma(t)) = \gamma(t + 2A)$$

provided that $\gamma(t)$ and $\gamma(t + 2A)$ are defined.

Thus if γ is defined originally on an interval $\gamma : [0, \lambda] \to G \times H$, γ may be extended to a geodesic $\widetilde{\gamma} : [0, 2\lambda] \to G \times H$ by choosing $q = \gamma(\lambda/2)$ and putting $\widetilde{\gamma}(t) = I_q I_p(\gamma(t - \lambda))$ for $t \in [\lambda, 2\lambda]$. It is then clear that γ may be defined on $(-\infty, \infty)$. Thus $(G \times H, \langle\langle\ ,\ \rangle\rangle)$ is geodesically complete. $\square$

Now Proposition 5.13 has the apparent defect that the existence of Lie groups $(G, \langle\ ,\ \rangle_1)$ equipped with bi-invariant Lorentzian metrics is assumed. It will now be shown how such Lie groups may be constructed by taking products of the form $(\mathbb{R} \times G, -dt^2 \oplus \langle\ ,\ \rangle)$ where $(G, \langle\ ,\ \rangle)$ is a Lie group equipped with a Riemannian bi-invariant metric.

The Lie group structure on $(\mathbb{R}, -dt^2)$ we will use is that induced by the usual addition of real numbers. Accordingly, we will write $(a, b) \mapsto a + b$ for the Lie group "multiplication" despite our use of the product notation above for the group operation. Here $\mathbb{R}$ is the analytic manifold determined by the chart $t : \mathbb{R} \to \mathbb{R}$, $t(r) = r$. Let $\partial/\partial t$ denote the corresponding coordinate vector field on $\mathbb{R}$. The left and right translation maps $L_a, R_a : \mathbb{R} \to \mathbb{R}$ are given by $L_a(r) = a + r$ and $R_a(r) = r + a$. It is easy to check that if $v = \lambda\, \partial/\partial t|_r \in T_r\mathbb{R}$, then $L_{a_*}v$ and $R_{a_*}v$ in $T_{a+r}(\mathbb{R})$ are given by $L_{a_*}v = R_{a_*}v = \lambda\, \partial/\partial t|_{a+r}$. Hence

$-dt^2(L_{a_*}v, L_{a_*}v) = -dt^2(R_{a_*}v, R_{a_*}v) = -\lambda^2 = -dt^2(v,v)$ so that $-dt^2$ is left and right invariant.

Let $(G, \langle\ ,\ \rangle)$ be a Lie group with a bi-invariant Riemannian metric. By the proof given in Proposition 5.13, G is a complete symmetric space. Also using (5.3) and (5.4), it is easily seen that the metric $\langle\langle\ ,\ \rangle\rangle = -dt^2 \oplus \langle\ ,\ \rangle$ is a bi-invariant Lorentzian metric for $\mathbb{R} \times G$. (Here, if $\xi_1 = (\lambda_1\, \partial/\partial t|_r\,, v_1)$ and $\xi_2 = (\lambda_2\, \partial/\partial t|_r\,, v_2)$ with $v_1, v_2 \in T_gG$, the inner product $\langle\langle \xi_1, \xi_2 \rangle\rangle = -\lambda_1\lambda_2 + \langle v_1, v_2 \rangle$.) Since $(G, \langle\ ,\ \rangle)$ is a complete Riemannian manifold, the product $(\mathbb{R} \times G, \langle\langle\ ,\ \rangle\rangle)$ is globally hyperbolic by Theorem 3.67. We have obtained

Theorem 5.14. *Let $(\mathbb{R}, -dt^2)$ be given the usual additive group structure and let $(G, \langle\ ,\ \rangle)$ be any Lie group equipped with a bi-invariant Riemannian metric. Then the product metric $\langle\langle\ ,\ \rangle\rangle = -dt^2 \oplus \langle\ ,\ \rangle$ is a bi-invariant Lorentzian metric for the product Lie group $\mathbb{R} \times G$. Thus $\big(\mathbb{R} \times G, \langle\langle\ ,\ \rangle\rangle\big)$ is a geodesically complete, globally hyperbolic space–time.*

Much research inspired by E. Cartan's work was done on Lie groups, homogeneous spaces, and symmetric spaces equipped with indefinite metrics before causality theory had assumed such a prominent role in general relativity. Thus most of this work was carried out not for Lorentzian metrics in particular but rather for general semi-Riemannian metrics of arbitrary signature. Rather than attempting to give an exhaustive list of references, we refer the reader to the bibliography in Wolf's (1974) text. Much of this research has been concerned with the problem of classifying all geodesically complete semi-Riemannian manifolds of constant curvature (the "space-form problem"). Two papers dealing with semi-Riemannian Lie theory have been written by Kulkarni (1978) and Nomizu (1979). Nomizu's paper deals specifically with Lorentzian metrics, considering the existence of constant curvature left-invariant Lorentzian metrics on a certain class of noncommutative Lie groups.

A more general treatment of the class of Lie groups which admit left invariant Lorentzian metrics of constant sectional curvature was then given in Barnet (1989). Also more recently, research in the Lie theory of semigroups, as reported in Hilgert, Hofmann, and Lawson (1989) or Lawson (1989), has sparked renewed interest in the causality and differential geometry of Lorentzian Lie groups on the part of the semigroup community. A representative research paper where causality and Lie semigroups are considered is Levichev and Levicheva (1992). Additional results for left invariant Lorentzian metrics on Lie groups of dimension three have recently been obtained by Cordero and Parker (1995b).

CHAPTER 6

COMPLETENESS AND EXTENDIBILITY

We mentioned in Chapter 1 that the Hopf–Rinow Theorem guarantees the equivalence of geodesic and metric completeness for arbitrary Riemannian manifolds. Further, either of these conditions implies the existence of minimal geodesics. That is, given any two points $p, q \in M$, there is a geodesic from p to q whose arc length realizes the metric distance from p to q. If M is compact, it also follows from the Hopf–Rinow Theorem that all Riemannian metrics for M are complete. In the noncompact case, Nomizu and Ozeki (1961) established that every noncompact smooth manifold admits a complete Riemannian metric. Extending their proof, Morrow (1970) showed that the complete Riemannian metrics for M are dense in the compact-open topology in the space of all Riemannian metrics for M [cf. Fegan and Millman (1978)].

In the first three sections of this chapter we compare and contrast these results with the theory of geodesic and metric completeness for arbitrary Lorentzian manifolds. In Section 6.1 a standard example is given to show that geodesic completeness does not imply the existence of maximal geodesic segments joining causally related points. Then we recall that the class of globally hyperbolic space–times possesses this useful property. In Section 6.2 we consider forms of completeness such as nonspacelike geodesic completeness, bounded acceleration completeness (b. a. completeness), and bundle completeness (b-completeness) that have been studied in singularity theory in general relativity [cf. Clarke and Schmidt (1977), Ellis and Schmidt (1977)]. We also state a corollary to Theorem 8 of Beem (1976a, p. 184) establishing the existence of nonspacelike complete metrics for all distinguishing space–times. In Section 6.3 we discuss Lorentzian metric completeness and the finite compactness condition.

In the last three sections of this chapter we discuss extensions and local extensions of space–times. Since extendibility is related to geodesic completeness, extendibility plays an important role in singularity theory in general relativity [cf. Clarke (1973, 1975, 1976), Hawking and Ellis (1973), Ellis and Schmidt (1977)]. In particular, one usually wants to avoid investigating space–times which are proper subsets of larger space–times since such proper subsets are always geodesically incomplete.

A space–time (M', g') is said to be an *extension* of a given space–time (M, g) if (M, g) may be isometrically embedded as a proper open subset of (M', g'). A space–time which has no extension is either said to be *inextendible* [cf. Hawking and Ellis (1973)] or *maximal* [cf. Sachs and Wu (1977a, p. 29)].

A local extension is an extension of a certain type of subset of a given space–time. In general, local inextendibility (i.e., the nonexistence of local extensions) implies global inextendibility. Since questions of extendibility naturally relate to the boundary of space–time, in Section 6.4 we briefly describe the Schmidt b-boundary and the Geroch–Kronheimer–Penrose causal boundary. In Section 6.5 two types of local extensions are defined and studied. If a Lorentzian manifold has no local extensions of either of these two types, it is shown to be inextendible. We also give a local extension of Minkowski space–time which shows that while b-completeness forces a space–time to be (globally) inextendible, b-completeness does *not* prevent a space–time from having local extensions.

In Section 6.6 local extensions are related to curvature singularities. For example, if (M, g) is an analytic space–time such that each timelike geodesic $\gamma : [0, a) \to M$ which is inextendible to $t = a$ is either complete (in the sense that $a = \infty$) or else corresponds to a curvature singularity, then (M, g) has no analytic local b-boundary extensions. Also, the a-boundary of Scott and Szekeres (1994) and its role in classifying singularities is discussed.

6.1 Existence of Maximal Geodesic Segments

The purpose of this section is twofold. First, we recall that for arbitrary Lorentzian manifolds, geodesic completeness does *not* imply the existence of

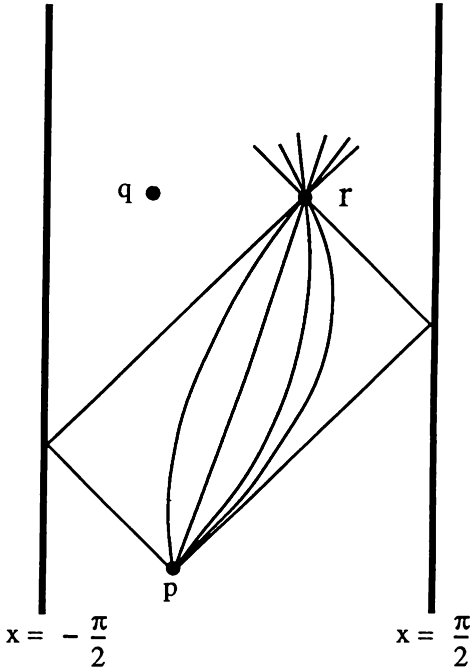


FIGURE 6.1. The universal cover $M = \{(x,t) : -\pi/2 < x < \pi/2\}$ of two-dimensional anti-de Sitter space is shown. The metric is given by $ds^2 = \sec^2 x\,(-dt^2 + dx^2)$. The points p and q are chronologically related in M, yet no maximal timelike geodesic in M joins p to q since all future directed timelike geodesics emanating from p are focused at r.

maximal geodesic segments joining causally related pairs of points. Second, we discuss the important and useful fact that distance realizing geodesics do exist for the class of globally hyperbolic space–times.

The universal covering manifold (M, g) of two-dimensional anti–de Sitter space provides an example that geodesic completeness does not imply that every pair $p, q \in M$ with $p \ll q$ may be joined by a timelike geodesic γ with $L(\gamma) = d(p, q)$ (cf. Figure 6.1). Recall that if γ is any future directed timelike curve from p to q with $L(\gamma) = d(p,q)$, then γ may be reparametrized to a timelike geodesic (cf. Theorem 4.13). Thus this same example shows that geodesically complete space–times exist which contain points $p \ll q$ such that $L(\gamma) < d(p,q)$ for all $\gamma \in \Omega_{p,q}$.

The space–time (M,g) may be represented by the strip $M = \{(x,t) \in \mathbb{R}^2 : -\pi/2 < x < \pi/2\}$ in $\mathbb{R}^2$ with the Lorentzian metric $ds^2 = \sec^2 x\,(-dt^2 + dx^2)$ [cf. Penrose (1972, p. 7)]. The points p and q in Figure 6.1 satisfy $p \ll q$, yet all future timelike geodesics emanating from p are focused again at the future timelike conjugate point r. Thus there is no timelike geodesic in M from p to q. Hence there is no maximal timelike geodesic or maximal timelike curve from p to q. Even more strikingly, it should be noted that an open set $U \subseteq I^+(p)$ of points of the type of q as in Figure 6.1 may be found with the property that none of the points of U are connected to p by any geodesic whatsoever, despite the geodesic completeness of this space–time.

We now consider which space–times do have the property that every pair of points $p, q \in M$ with $q \in J^+(p)$ may be joined by a distance realizing geodesic. If $M = \mathbb{R}^2 - \{(0,0)\}$ with the Lorentzian metric $ds^2 = dx^2 - dy^2$, then $p = (0,-1)$ and $q = (0,1)$ are points in M with $d(p,q) = 2 > 0$ which cannot be joined by a maximal timelike geodesic. [The desired geodesic would have to be the curve $\gamma(t) = (0,t)$, $-1 \leq t \leq 1$, which passes through the deleted point $(0,0)$.] On the other hand, this space–time is chronological, strongly causal, and stably causal. Thus it is reasonable to restrict our attention to the class of globally hyperbolic space–times. For these space–times, Avez (1963) and Seifert (1967) have shown that given any $p, q \in M$ with $p \leqslant q$, there is a geodesic from p to q which maximizes arc length among all nonspacelike future directed curves from p to q (cf. Theorem 3.18). In the language of Definition 4.10, this may be stated as follows.

Theorem 6.1. *Let (M,g) be globally hyperbolic. Then given any $p, q \in M$ with $q \in J^+(p)$, there is a maximal geodesic segment $\gamma \in \Omega_{p,q}$, i.e., a future directed nonspacelike geodesic γ from p to q with $L(\gamma) = d(p,q)$.*

We sketch Seifert's (1967, Theorem 1) proof of this result [cf. Penrose (1972, Chapter 6)]. Since (M,g) is globally hyperbolic, it may be shown that if

$p \leqslant q$, the nonspacelike path space $\Omega_{p,q}$ is compact. On the other hand, since (M, g) is strongly causal, the arc-length functional $L : \Omega_{p,q} \to \mathbb{R}$ is upper semicontinuous in the C^0 topology (cf. Section 3.3). Thus there exists a curve $\gamma_0 \in \Omega_{p,q}$ with $L(\gamma_0) = \sup\{L(\gamma) : \gamma \in \Omega_{p,q}\}$. It follows from the variational theory of arc length that if γ_0 is not a reparametrization of a smooth geodesic, a curve $\sigma \in \Omega_{p,q}$ with $L(\sigma) > L(\gamma)$ may be constructed, in contradiction. Alternatively, if $L(\gamma_0) = \sup\{L(\gamma) : \gamma \in \Omega_{p,q}\}$, then $L(\gamma_0) = d(p, q)$ by the definition of Lorentzian distance. Hence Theorem 4.13 implies that γ_0 is, up to reparametrization, a smooth geodesic.

In the case that $p \ll q$, the maximal curve γ_0 may also be constructed using the results of Section 3.3. Let $h : M \to \mathbb{R}$ be a globally hyperbolic time function for (M, g). Choose t_0 with $h(p) < t_0 < h(q)$. Then $K = J^+(p) \cap J^-(q) \cap h^{-1}(t_0)$ is compact, and any nonspacelike curve from p to q intersects K. By definition of Lorentzian distance, we may find a curve $\gamma_n \in \Omega_{p,q}$ with

$$d(p, q) \geq L(\gamma_n) \geq d(p, q) - \frac{1}{n} \tag{6.1}$$

for each positive integer n. Let $r_n \in \gamma_n \cap K$. Since K is compact, a subsequence $\{r_{n(j)}\}$ converges to $r \in K$. By Corollary 3.32, there is a nonspacelike limit curve γ_0 of the sequence $\{\gamma_{n(j)}\}$ passing through r and joining p to q. Since (M, g) is strongly causal, a subsequence of $\{\gamma_{n(j)}\}$ converges to γ_0 in the C^0 topology by Proposition 3.34. Using Remark 3.35 and condition (6.1), we obtain $L(\gamma_0) \geq d(p, q)$. Hence by the definition of distance, $L(\gamma_0) = d(p, q)$, and γ_0 may be reparametrized to a smooth geodesic by Theorem 4.13. If $p \leqslant q$ and $d(p, q) = 0$, we already know that there is a maximal null geodesic segment from p to q by Corollary 4.14.

In connection with Theorem 6.1, it should be noted that global hyperbolicity is not a necessary condition for the existence of maximal geodesic segments joining all pairs of causally related points. For let $M = \{(x, y) \in \mathbb{R}^2 : 0 < x < 10,\ 0 < y < 10\}$ be equipped with the Lorentzian metric it inherits as an open subset of Minkowski space. Since the geodesics in M are just Euclidean straight line segments, it is readily seen that maximal geodesics exist joining any pair of causally related points. However, if $p = (1, 1)$ and $q = (1, 9)$, then

$J^+(p) \cap J^-(q)$ is noncompact. Hence this space–time, while strongly causal, fails to be globally hyperbolic.

6.2 Geodesic Completeness

We showed in Theorem 4.9 that for space–times which are strongly causal, the Lorentzian distance function may be used to construct a subbasis for the given manifold topology. Nonetheless, the sets $\{q \in M : d(p,q) < R\}$ fail to form a basis for the given manifold topology. Thus geodesic completeness rather than metric completeness of space–times has usually been considered in general relativity.

Let (M, g) be an arbitrary Lorentzian manifold.

Definition 6.2. (*Complete Geodesic*) A geodesic c in (M, g) with affine parameter t is said to be *complete* if the geodesic can be extended to be defined for $-\infty < t < \infty$. A past and future inextendible geodesic is said to be *incomplete* if it cannot be extended to arbitrarily large positive and negative values of an affine parameter. *Future* or *past incomplete* geodesics may be defined similarly.

An affine parameter t for the curve c is a parametrization such that $c(t)$ satisfies the geodesic differential equation $\nabla_{c'} c'(t) = 0$ for all t [cf. Kobayashi and Nomizu (1963, p. 138)]. It is necessary to use the concept of an affine parameter since null geodesics, which have zero arc length, cannot be parametrized by arc length. If s and t are two affine parameters for c, it follows from the geodesic differential equations that there exist constants $\alpha, \beta \in \mathbb{R}$ such that $s(t) = \alpha t + \beta$ for all t in the domain of c. Hence completeness or incompleteness as defined in Definition 6.2 is independent of the choice of affine parameter. In particular, if c is an inextendible timelike geodesic parametrized by arc length (i.e., $g(c'(t), c'(t)) = -1$ for all t in the domain of c), then c is incomplete if $L(c) < \infty$. Even if $L(c) = \infty$, it may happen that c is incomplete. This occurs, for example, when the domain of c is of the form (a, ∞), where $a > -\infty$.

Certain exact solutions to the Einstein equations in general relativity, like the extended Schwarzschild solution, contain nonspacelike geodesics which become incomplete upon running into black holes. Even though the existence of incomplete, inextendible, nonspacelike geodesics does not force a space–time to contain a black hole, these examples suggest that nonspacelike geodesic incompleteness might be used as a first order test for "singular space–times" [cf. Hawking and Ellis (1973, Chapter 8), Clarke and Schmidt (1977), Ellis and Schmidt (1977)]. Thus it is standard to make the following definitions in general relativity. Recall that a geodesic is said to be inextendible if it is both past and future inextendible.

Definition 6.3. (*Geodesically Complete*) A space–time (M, g) is said to be *timelike* (respectively, *null, nonspacelike, spacelike*) *geodesically complete* if all timelike (respectively, null, nonspacelike, spacelike) inextendible geodesics are complete. The space–time (M, g) is said to be *geodesically complete* if all inextendible geodesics are complete. Also, (M, g) is said to be *timelike* (respectively, *null, nonspacelike, spacelike*) *geodesically incomplete* if some timelike (respectively, null, nonspacelike, spacelike) geodesic is incomplete. A nonspacelike incomplete space–time is said to be a *geodesically singular space–time*.

It was once hoped that timelike geodesic completeness might imply null geodesic completeness, etc. However, Kundt (1963) gave an example of a space–time that is timelike and null geodesically complete but not spacelike complete. Then Geroch (1968b, p. 531) gave an example of a space–time conformal to Minkowski two-space and thus globally hyperbolic which is timelike incomplete but null and spacelike complete. Also Geroch remarked that modifications of Kundt's and his examples gave space–times that were (1) incomplete in any two ways but complete in the third way, (2) spacelike incomplete but null and timelike complete, and (3) timelike incomplete but spacelike and null complete. Then Beem (1976c) gave an example of a globally hyperbolic space–time that was null incomplete but spacelike and timelike complete. These results may be summarized as follows.

Theorem 6.4. *Timelike geodesic completeness, null geodesic completeness, and spacelike geodesic completeness are all logically inequivalent.*

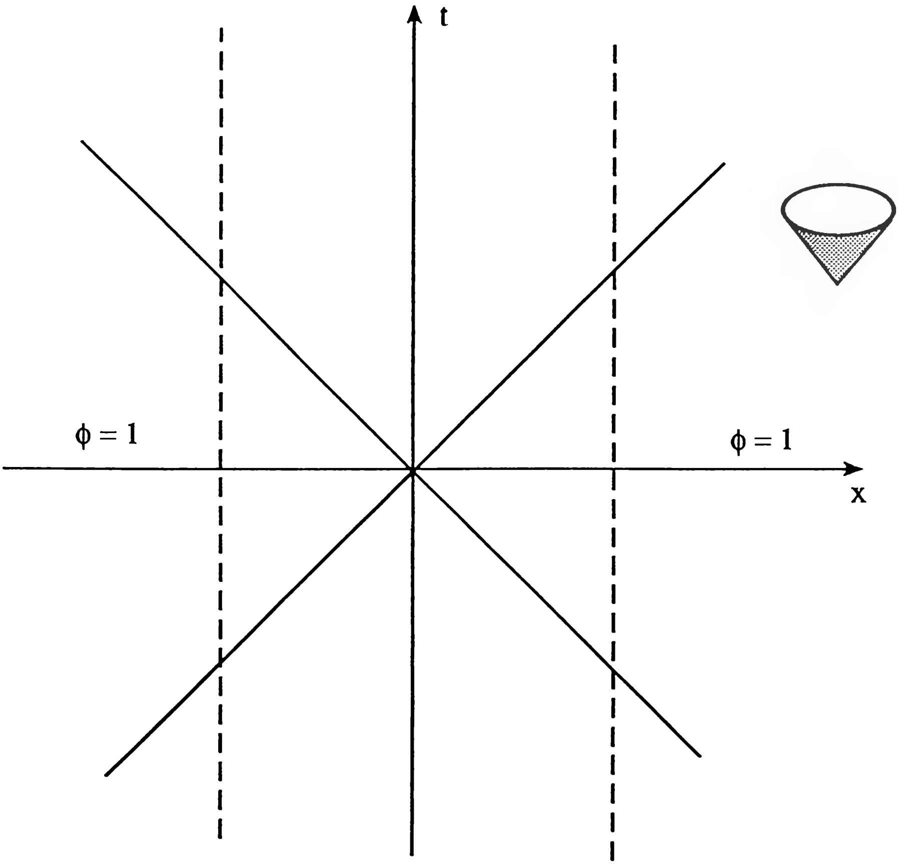


FIGURE 6.2. Shown is Geroch's example of a space–time globally conformal to Minkowski two-space, which is null and spacelike geodesically complete but timelike geodesically incomplete. Here the positive t axis may be parametrized to be an incomplete timelike geodesic since $\phi(0,t) \to 0$ like t^{-4} as $t \to \infty$.

In order to illustrate the constructions used in the proof of Theorem 6.4, we now describe Geroch's example of a space–time which is null and spacelike complete but timelike incomplete. Let $(\mathbb{R}^2, g_1)$ be Minkowski two-space with global coordinates (x,t) and the usual Lorentzian metric $g_1 = ds^2 = dx^2 - dt^2$.

Conformally change the metric g_1 to a new metric $g = \phi g_1$ for $\mathbb{R}^2$, where $\phi : \mathbb{R}^2 \to (0, \infty)$ is a smooth function with the following properties (cf. Figure 6.2):

(1) $\phi(x, t) = 1$ if $x \leq -1$ or $x \geq 1$;
(2) $\phi(x, t) = \phi(-x, t)$ for all $(x, t) \in \mathbb{R}^2$; and
(3) On the t axis, $\phi(0, t)$ goes to zero like t^{-4} as $t \to \infty$.

Since g is conformal to g_1, the space–time $(\mathbb{R}^2, g)$ is globally hyperbolic, and null geodesics still have as images straight lines making angles of 45° with the positive or negative x axis. By property (2), the reflection $F(x, t) = (-x, t)$ is an isometry of $(\mathbb{R}^2, g)$. Since the fixed point set of an isometry is totally geodesic, the t axis may be parametrized as a timelike geodesic. By condition (3), this geodesic is incomplete as $t \to \infty$. Thus $(\mathbb{R}^2, g)$ is timelike incomplete. But every null or spacelike geodesic which enters the region $-1 \leq x \leq 1$ eventually leaves and then remains outside this region. Thus condition (1) implies that $(\mathbb{R}^2, g)$ is null and spacelike complete.

We now consider the converse problem of constructing geodesically complete Lorentzian metrics for paracompact smooth manifolds. In order to preserve the causal structure of the given space–time, we restrict our attention to global conformal changes rather than arbitrary metric deformations.

For Riemannian metrics, Nomizu and Ozeki (1961) showed that an arbitrary metric can be made complete by a global conformal change. On the other hand, space–times exist with the property that no global conformal factor will make these space–times nonspacelike geodesically complete. A two-dimensional example with this property has been given by Misner (1967). In this example there are inextendible null geodesics which are future incomplete and future trapped in a compact set [cf. Hawking and Ellis (1973, pp. 171–172)]. Any conformal change of this example will leave these null geodesics pointwise fixed and future incomplete. Thus one may not establish an analogue of the Nomizu and Ozeki result for arbitrary space–times.

However, the existence of nonspacelike complete Lorentzian metrics has been shown for space–times satisfying certain causality conditions. Seifert (1971, p. 258) has shown that if (M, g) is stably causal, then M is conformal to a space–time with all future directed (or all past directed) nonspacelike geodesics complete. Also Clarke (1971) has shown that a strongly causal space–time may be made null geodesically complete by a conformal factor. Beem (1976a) studied space–times with the property that for each compact subset K of M, no future inextendible nonspacelike curve is future imprisoned in K. (Recall that the nonspacelike curve γ is said to be future imprisoned in K if there exists $t_0 \in \mathbb{R}$ such that $\gamma(t) \in K$ for all $t \geq t_0$.) If (M, g) is a causal space–time satisfying this condition, then there exists a conformal factor $\Omega : M \to (0, \infty)$ such that $(M, \Omega g)$ is null and timelike geodesically complete [Beem (1976a, p. 184, Theorem 8)]. This imprisonment condition is satisfied if (M, g) is stably causal, strongly causal, or distinguishing. Hence we may state the following result.

Theorem 6.5. *If (M, g) is distinguishing, strongly causal, stably causal, or globally hyperbolic, then there exists a smooth conformal factor $\Omega : M \to (0, \infty)$ such that the space–time $(M, \Omega g)$ is timelike and null geodesically complete.*

It is an open question as to whether Theorem 6.5 can be strengthened to include spacelike geodesic completeness as well (cf. Corollary 3.46 for space–times homeomorphic to $\mathbb{R}^2$).

Suppose that a space–time is defined to be nonsingular if it is geodesically complete. Then "no regions have been deleted from the space–time manifold" of a nonsingular space–time [Geroch (1968b, Property 1)]. But Geroch (1968b, Property 2) suggested a second condition that nonsingular space–times should satisfy, namely, "observers who follow 'reasonable' (in some sense) world lines should have an infinite total proper time." Here a "world line" is a timelike curve in (M, g). Then Geroch (1968b, pp. 534–540) constructed a geodesically complete space–time which contains a smooth timelike *curve* of bounded acceleration but having finite length. Thus this example fails to satisfy Geroch's Property 2 even though all timelike *geodesics* have infinite length by the geodesic completeness.

Accordingly, in addition to geodesically incomplete space–times, further kinds of singular space–times have been studied in general relativity. In the rest of this section, we will discuss two of these additional types of completeness, b.a. completeness (bounded acceleration completeness) and b-completeness (bundle completeness).

The concept of b.a. completeness stems from the preceding example of Geroch. For the purpose of stating Definition 6.6, we recall than any C^2 timelike curve may be reparametrized to a C^2 timelike curve $\gamma : J \to M$ with $g(\gamma'(t), \gamma'(t)) = -1$ for all $t \in J$.

Definition 6.6. (*Bounded Acceleration*) A C^2 timelike curve $\gamma : J \to M$ with $g(\gamma'(t), \gamma'(t)) = -1$ for all $t \in J$ is said to have *bounded acceleration* if there exists a constant $B > 0$ such that $\left|g\left(\nabla_{\gamma'}\gamma'(t), \nabla_{\gamma'}\gamma'(t)\right)\right| \leq B$ for all $t \in J$.

Here ∇ is the unique torsion free connection for M defined by the metric g [cf. Section 2.2, equations (2.16) and (2.17)]. In particular, if γ is a geodesic, then γ has zero and hence bounded acceleration. The requirement that γ be C^2 makes it possible to calculate $\nabla_{\gamma'}\gamma'$.

Definition 6.7. (*b.a. complete space–time*) A space–time (M, g) is said to be *b.a. complete* if all future (respectively, past) directed, future (respectively, past) inextendible, unit speed, C^2 timelike curves with bounded acceleration have infinite length. If there exists a future (or past) directed, future (or past) inextendible, unit speed, C^2 timelike curve with bounded acceleration but finite length, then (M, g) is said to be *b.a. incomplete*.

Geroch's example (1968b, pp. 534–540) shows that geodesic completeness does not imply b.a. completeness. Furthermore, Beem (1976c, p. 509) has given an example to show that even for globally hyperbolic space–times, geodesic completeness does not imply b.a. completeness. Trivially, b.a. completeness implies timelike geodesic completeness. On the other hand, the example of Geroch given in Figure 6.2 may be modified by changing the sign of the metric tensor to show that b.a. completeness does not imply spacelike geodesic completeness.

A stronger form of completeness, *b-completeness*, does imply geodesic completeness and hence overcomes this last objection to b.a. completeness. The concept of b-completeness, which was first studied for Lorentzian manifolds by Schmidt (1971), is intuitively defined as follows [cf. Hawking and Ellis (1973, p. 259)]. First, the concept of an affine parameter is extended from geodesics to all C^1 curves. Then a space–time is said to be b-complete if every C^1 curve of finite length in such a parameter has an endpoint.

We now give a brief discussion of b-completeness. First, it is necessary to discuss the concept of a *generalized affine parameter* for any C^1 curve $\gamma : J \to M$. Recall that a smooth vector field V along γ is a smooth map $V : J \to TM$ such that $V(t) \in T_{\gamma(t)}M$ for all $t \in J$. Such a smooth vector field V along γ is said to be a parallel field along γ if V satisfies the differential equation $\nabla_{\gamma'}V(t) = 0$ for all $t \in J$ (cf. Chapter 2).

A generalized affine parameter $\mu = \mu(\gamma, E_1, E_2, \ldots, E_n)$ may be constructed for $\gamma : J \to M$ as follows. Choosing any $t_0 \in J$, let $\{e_1, e_2, \ldots, e_n\}$ be any basis for $T_{\gamma(t_0)}M$. Let E_i be the unique parallel field along γ with $E_i(t_0) = e_i$ for $1 \leq i \leq n$. Then $\{E_1(t), E_2(t), \ldots, E_n(t)\}$ forms a basis for $T_{\gamma(t)}M$ for each $t \in J$. We may thus write $\gamma'(t) = \sum_{i=1}^n V^i(t)E_i(t)$ with $V^i : J \to \mathbb{R}$ for $1 \leq i \leq n$. Then the generalized affine parameter $\mu = \mu(\gamma, E_1, \ldots, E_n)$ is given by

$$\mu(t) = \int_{t_0}^{t} \sqrt{\sum_{i=1}^{n} [V^i(s)]^2}\, ds, \qquad t \in J.$$

The assumption that γ is C^1 is necessary in order to obtain the vector fields $\{E_1, E_2, \ldots, E_n\}$ by parallel translation. It may be checked that γ has finite arc length in the generalized affine parameter $\mu = \mu(\gamma, E_1, \ldots, E_n)$ if and only if γ has finite arc length in any other generalized affine parameter $\mu = \mu(\gamma, \overline{E}_1, \ldots, \overline{E}_n)$ calculated from any other basis $\{\overline{E}_i\}_{i=1}^n$ for $TM|_\gamma$ obtained by parallel translation along γ [cf. Hawking and Ellis (1973, p. 259)]. Hence the concept of finite arc length with respect to a generalized affine parameter is independent of the particular choice of generalized affine parameter. It thus makes sense to make the following definition.

Definition 6.8. (*b-complete space–time*) The space–time (M, g) is said to be *b-complete* if every C^1 curve of finite arc length as measured by a generalized affine parameter has an endpoint in M.

Suppose $\gamma : J \to M$ is any smooth geodesic. Taking $E_1(t) = \gamma'(t)$ in the above construction, for any choice of $E_2, \ldots, E_n$ we have generalized affine parameter $\mu(\gamma, E_1, E_2, \ldots, E_n)(t) = t$. Hence b-completeness implies geodesic completeness. It is also known that b-completeness implies b.a. completeness. Geroch's example (cf. Figure 6.2) with the sign of the metric tensor changed shows that there are globally hyperbolic space–times which are b.a. complete but not b-complete. Thus b.a. completeness does not imply b-completeness.

6.3 Metric Completeness

The Hopf–Rinow Theorem for Riemannian manifolds (N, g_0) implies that the following are equivalent:

(1) N with the Riemannian distance function $d_0 : N \times N \to [0, \infty)$ is a complete metric space, i.e., all Cauchy sequences converge.

(2) (N, d_0) is finitely compact, i.e., all d_0-bounded sets have compact closure.

(3) (N, g_0) is geodesically complete.

Here a set K in a Riemannian manifold (N, g_0) is said to be *bounded* if $\sup\{d_0(p, q) : p, q \in K\} < \infty$. By the triangle inequality, this is equivalent to the condition that K be contained inside a closed metric ball of finite radius.

In Section 6.2 we considered the geodesic completeness of Lorentzian manifolds. In this section we shall consider Lorentzian analogues of conditions (1) and (2) above. From the very definition of Lorentzian distance [i.e., $d(p, q) = 0$ if $q \notin J^+(p)$], it is clear that attention should be restricted to timelike Cauchy sequences.

Busemann (1967) studied general Hausdorff spaces having a partial ordering with properties similar to those of the chronological partial ordering $p \ll q$ of a space–time. Also, Busemann supposed that these spaces, which he called *timelike spaces*, were equipped with a function which behaves just like the Lorentzian distance function of a chronological space–time restricted to the set $\{(p, q) \in M \times M : p \leqslant q\}$. For this class of nondifferentiable spaces, Busemann observed that the length of continuous curves could be defined and, moreover,

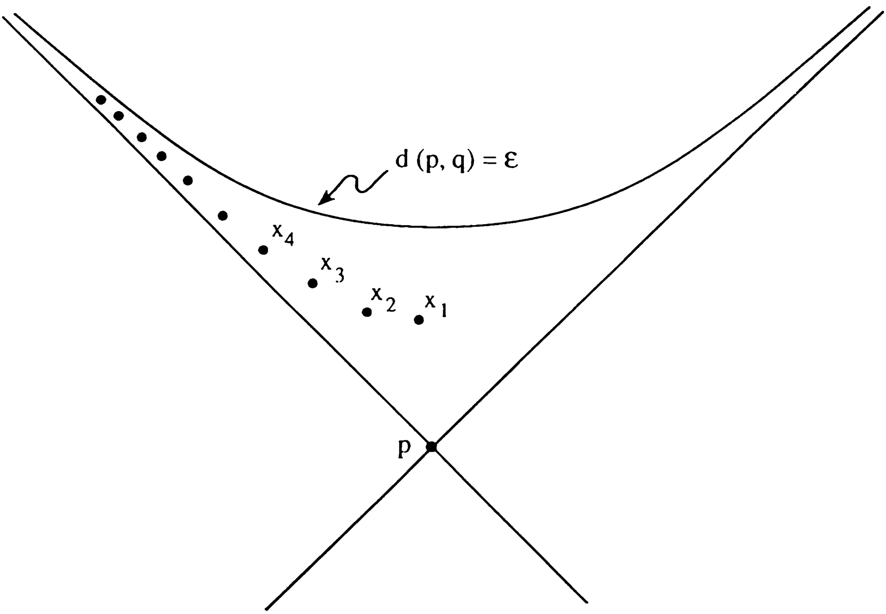


FIGURE 6.3. Shown is a sequence $\{x_n\}$ in Minkowski two-space $(\mathbb{R}^2, ds^2 = dx^2 - dy^2)$ with $x_n \gg p$ for all n, $d(p, x_n) \to 0$ as $n \to \infty$, but such that $\{x_n\}$ has no point of accumulation.

the length functional is upper semicontinuous in a topology of uniform convergence [cf. Busemann (1967, p. 10)]. Busemann's aim in studying timelike spaces was to develop a geometric theory for indefinite metrics analogous to the theory of metric G-spaces [cf. Busemann (1955)]. In particular, Busemann studied finite compactness and metric completeness for timelike spaces in the spirit of (1) and (2) of the Hopf–Rinow Theorem.

Beem (1976b) observed that Busemann's definitions of finite compactness and metric completeness for timelike G-spaces may be adapted to causal space–times. First, timelike Cauchy completeness may be defined for causal space–times as follows.

Definition 6.9. (*Timelike Cauchy Complete*) The causal space–time (M, g) is said to be *timelike Cauchy complete* if any sequence $\{x_n\}$ of points with $x_n \ll x_{n+m}$ for $n, m = 1, 2, 3, \ldots$ and $d(x_n, x_{n+m}) \leq B_n$ [or else $x_{n+m} \ll x_n$ and $d(x_{n+m}, x_n) \leq B_n$] for all $m \geq 0$, where $B_n \to 0$ as $n \to \infty$, is a convergent sequence.

For Riemannian manifolds, finite compactness may be defined by requiring that all closed metric balls be compact. On the other hand, we have noted above (cf. Figure 4.4) that the subsets $\{q \in J^+(p) : d(p, q) \leq \epsilon\}$ of a space–time are generally noncompact. Thus the Riemannian definition must be modified. One possibility is the following [cf. Busemann (1967, p. 22)].

Definition 6.10. (*Finitely Compact*) The causal space–time (M, g) is said to be *finitely compact* if for each fixed constant $B > 0$ and each sequence of points $\{x_n\}$ with either $p \ll q \leqslant x_n$ and $d(p, x_n) \leq B$ for all n, or $x_n \leqslant q \ll p$ and $d(x_n, p) \leq B$ for all n, there is a point of accumulation of $\{x_n\}$ in M.

It may be seen that without requiring $p \ll q \leqslant x_n$ (or $x_n \leqslant q \ll p$) for some $q \in M$ in Definition 6.10, Minkowski space–time fails to be finitely compact (cf. Figure 6.3).

For globally hyperbolic space–times, a characterization of finite compactness more reminiscent of condition (2) above for Riemannian manifolds may be given.

Lemma 6.11. *Let (M, g) be globally hyperbolic. Then (M, g) is finitely compact iff for each real constant $B > 0$, the set $\{x \in M : p \ll q \leqslant x,\ d(p, x) \leq B\}$ is compact for any $p, q \in M$ with $q \in I^+(p)$ and the set $\{x \in M : x \leqslant q \ll p,\ d(x, p) \leq B\}$ is compact for any $p, q \in M$ with $p \in I^+(q)$.*

Proof. This now follows easily because the sets $J^+(q)$ are closed and the Lorentzian distance function is continuous since (M, g) is globally hyperbolic. □

Minkowski space–time is both timelike Cauchy complete and finitely compact. More generally, it may be shown that these concepts are equivalent for all globally hyperbolic space–times [cf. Beem (1976b, pp. 343–344)].

Theorem 6.12. *If (M, g) is globally hyperbolic, then (M, g) is finitely compact iff (M, g) is timelike Cauchy complete. Also, if (M, g) is globally hyperbolic and nonspacelike geodesically complete, then (M, g) is finitely compact and timelike Cauchy complete.*

Remark 6.13. Even for the class of globally hyperbolic space–times, finite compactness, or equivalently, timelike Cauchy completeness, does *not* imply timelike geodesic completeness. Indeed, Geroch's example given in Figure 6.2 is a timelike geodesically incomplete, globally hyperbolic space–time which is finitely compact.

At times, it is important to consider the completeness of a submanifold as well as that of the given manifold. Conditions which are sufficient to guarantee the completeness of submanifolds of Lorentzian manifolds are, in general, more complicated than corresponding conditions for submanifolds of Riemannian manifolds [cf. Beem and Ehrlich (1985a,b), Harris (1987, 1988a,b, 1994)]. If (H, h) is a complete Riemannian manifold and $F : M \to H$ is an embedding of M with $F(M)$ a closed subset of H, then M is a complete Riemannian manifold using the induced metric. The converse of this result is false. For example, let the curve $c : (0, +\infty) \to \mathbb{R}^2$ be given by $c(t) = (t, \sin(1/t))$. This curve has an image which is a complete submanifold of the usual Euclidean plane, but this image clearly fails to be a closed subset of $\mathbb{R}^2$.

Unlike the Riemannian case, closed embedded submanifolds of Lorentzian manifolds need not be complete. To see this, it suffices to take a curve in the Minkowski plane which is asymptotic to a null line quickly enough in at least one direction along the curve. For example, the following map [cf. Harris (1988a)]

$$F(x) = \left(x, \int_0^{|x|} [1 - e^{-t}]^{\frac{1}{2}} \, dt \right)$$

is an embedding of $\mathbb{R}^1$ as an incomplete closed spacelike submanifold of the (x, y) plane with the usual Minkowski metric $\eta = dx^2 - dy^2$. Harris (1988a) has studied the completeness of embedded and immersed spacelike hypersurfaces of Minkowski space. Among other things, he has shown that if such an immersed hypersurface in $(n+1)$-dimensional Minkowski space (i.e., L^{n+1}) is complete, then its image must be a closed achronal subset which is diffeomorphic to $\mathbb{R}^n$ and it must be embedded as a graph of a function defined on a spacelike hyperplane. A more detailed resume of Harris's work on spacelike completeness together with additional references is given in Harris (1993, 1994).

Separate investigations had earlier been conducted in the case of *constant mean curvature* spacelike hypersurfaces S of L^{n+1} which are closed subsets of L^{n+1} in the Euclidean topology. For this class of space–times, Cheng and Yau (1976) showed that the condition of constant mean curvature H_0 implies that S is complete in the induced Riemannian metric. (A complete treatment of the issue of the achronality en route to the proof of completeness is given in Harris (1988a, pp. 112, 118). Also, Harris (1988a, p. 118) showed that the hypothesis of bounded principal curvatures may be substituted for constant mean curvature.) Further, Cheng and Yau (1976) showed that constant mean curvature H_0 implies that the length of the second fundamental form is bounded by $n\,|H_0|$ and also that S has nonpositive Ricci curvature. In particular, in the case of a maximal (i.e., $H_0 = 0$) spacelike hypersurface, this estimate shows that the second fundamental form is trivial. Hence these results of Cheng and Yau (1976) combine with earlier work of Calabi (1968) for $n = 3$ to yield the result that the only maximal spacelike hypersurface which is a closed subset of Minkowski space is a linear hyperplane. Nishikawa (1984) investigated the more general case of a locally symmetric target manifold satisfying the timelike convergence condition and a condition that the sectional curvature of all nondegenerate two-planes spanned by a pair of spacelike vectors be nonnegative. Nishikawa showed that a complete maximal spacelike hypersurface in such a target space–time would be totally geodesic.

In the case of nonzero constant mean curvature, Goddard (1977b) carried out perturbation calculations for hyperboloids in Minkowski space (and also for appropriate submanifolds of de Sitter space–time) which suggested that perhaps all entire, constant mean curvature, spacelike hypersurfaces of Minkowski space should be hyperboloids. This issue was thoroughly investigated by Treibergs (1982), who took the starting point that in the case of positive mean curvature, an entire spacelike hypersurface could be realized as the graph of a *convex* function. Now for an entire convex function f whose graph is a spacelike hypersurface, Treibergs (1982, p. 51) defines the *projective boundary values of f at infinity*, V_f, as

$$V_f(x) = \lim_{r \to +\infty} \frac{f(rx)}{r}.$$

Treibergs defines two entire, constant mean curvature, spacelike hypersurfaces to be *equivalent* if they have the same projective boundary values at infinity. Treibergs then proves that the set of such equivalence classes coincides with convex homogeneous functions whose gradient has norm one whenever defined. Treibergs further shows that constant mean curvature spacelike hypersurfaces with given projective boundary data at infinity are highly nonunique, because arbitrary finite perturbations of the given light cone (at infinity) may be made producing different $f(x)$, hence different spacelike hypersurfaces, each strongly asymptotic to its own perturbed light cone, yet all having the same projective boundary behavior $V(x)$ at infinity. Thus the geometry for entire, constant (but nonzero) mean curvature, spacelike hypersurfaces turns out to be more complicated than the maximal ($H_0 = 0$) case.

6.4 Ideal Boundaries

In this section we give brief descriptions of the b-boundary and the causal boundary for a space–time. Further details may be found in Hawking and Ellis (1973, Sections 6.8 and 8.3) or Dodson (1978).

The *b-boundary* of a space–time (M, g) will be denoted by $\partial_b M$. This boundary is formed by defining a certain positive definite metric on the bundle of linear frames $L(M)$ over M, taking the Cauchy completion of $L(M)$, and then using the newly formed ideal points of $L(M)$ to obtain ideal points of M. The b-boundary is particularly useful in telling whether or not some points have been removed from the space–time. Somewhat unfortunately, the b-boundary often consists of just a single point [cf. Bosshard (1976), Johnson (1977)]. This boundary is not invariant under conformal changes and also is not directly related to the causal structure of (M, g). A discussion of the merits and demerits of the b-boundary and geodesic incompleteness may be found in the review article of Tipler, Clarke, and Ellis (1980).

Recall that a curve $\gamma : [0, a) \to M$ is said to be *b-incomplete* if it has finite generalized affine parameter (cf. Section 6.2). Any curve $\gamma : [0, a) \to M$ which is both b-incomplete and inextendible to $t = a$ defines a point of $\partial_b M$ corresponding to $\gamma(a)$. In Minkowski space–time, generalized affine parameter values along a curve can be made to correspond to Euclidean arc length. Thus Minkowski space–time has an empty b-boundary, and each b-incomplete curve in Minkowski space–time has an endpoint in the space–time.

The *causal boundary* of a space–time (M, g) will be denoted by $\partial_c M$. This boundary is constructed using the causal structure of the space–time. Thus it is invariant under conformal changes. We will only be interested in considering this boundary for strongly causal space–times.

The causal boundary is formed using indecomposable past (respectively, future) sets which do not correspond to the past (respectively, future) of any point of M. A *past* (respectively, *future*) *set* A is a subset of M such that $I^-(A) \subseteq A$ (respectively, $I^+(A) \subseteq A$). The open past (respectively, future) sets are characterized by $I^-(A) = A$ (respectively, $I^+(A) = A$). An *indecomposable past set* (IP) is an open past set that cannot be written as a union of two proper subsets both of which are open past sets. An *indecomposable future set* (IF) is defined dually.

A *terminal indecomposable past set* (TIP) is a subset A of M such that

(1) A is an indecomposable past set, and
(2) A is not the chronological past of any point $p \in M$.

A *terminal indecomposable future set* (TIF) is defined dually. The causal boundary $\partial_c M$ is formed using TIP's and TIF's after making certain identifications which will be described below [cf. Hawking and Ellis (1973, pp. 218–221)]. These identifications allow the topology of M to be extended to $M^* = M \cup \partial_c M$ in such a way that the causal completion of M is Hausdorff, provided that (M, g) satisfies certain more restrictive causality conditions such as stable causality.

The use of TIP's and TIF's to represent ideal points of the causal boundary of (M, g) is illustrated in Figure 6.4.

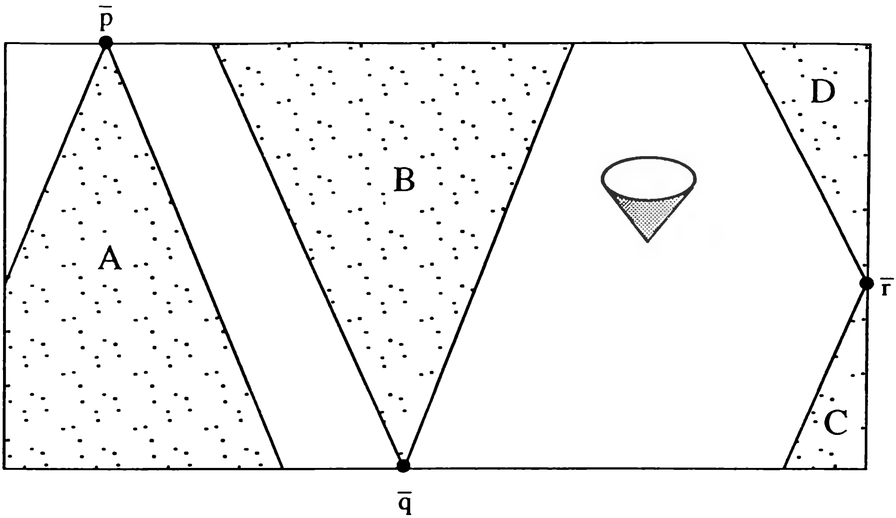


FIGURE 6.4. The ideal point $\overline{p}$ in $\partial_c M$ is represented by the terminal indecomposable past set A, and the ideal point $\overline{q}$ is represented by the terminal indecomposable future set B. The point $\overline{r}$ is represented by both the set C, which is a TIP, and the set D, which is a TIF.

We now show that a TIP may be represented as the chronological past of a future inextendible timelike curve. This result is due to Geroch, Kronheimer, and Penrose (1972, p. 551).

Proposition 6.14. *A subset W of the strongly causal space–time (M, g) is a TIP iff there exists a future directed and future inextendible timelike curve γ such that $W = I^-(\gamma)$.*

Proof. Assume there is a future inextendible timelike curve γ with $W = I^-(\gamma)$. Using the strong causality of (M, g), it follows that if W is an IP, then W is a TIP. To show W is an IP, assume that $W = U_1 \cup U_2$ for nonempty open past sets U_1 and U_2 such that neither is a subset of the other. Choose $r_1 \in U_1 - U_2$ and $r_2 \in U_2 - U_1$. There must exist points $r_i' \in \gamma$ such that $r_i \in I^-(r_i')$ for $i = 1, 2$ because $U_1 \cup U_2 = I^-(\gamma)$. However, whichever U_i contains the futuremost of r_1' and r_2' must then contain all four of r_1, r_2, r_1', and r_2'. This contradicts either the definition of r_1 or of r_2.

On the other hand, assume that W is a TIP. If p is any point of W, then $W = [W \cap I^+(p)] \cup [W - I^+(p)]$ and thus $W = I^-(W \cap I^+(p)) \cup I^-(W - I^+(p))$ since W is a past set. Since W is an IP, either $W = I^-(W \cap I^+(p))$ or $W = I^-(W - I^+(p))$. Consequently, as $p \notin I^-(W - I^+(p))$, we have $W = I^-(W \cap I^+(p))$. Thus, given any $q \neq p$ in W, there must be some point r in W which is in the chronological future of both p and q. Inductively, for each finite subset of W, there exists some point of W in the chronological future of each point of the subset. Now choose a sequence of points $\{p_n\}$ which forms a countable dense subset of W. We will define a second sequence $\{q_n\}$ inductively. Let q_0 be a point of W in the chronological future of p_0. If q_i for $i = 1, 2, \ldots, k-1$ has been defined, then choose q_k to be a point of W in the chronological future of p_k and of q_i for $i = 1, 2, \ldots, k-1$.

Finally let γ be any future directed timelike curve which begins at q_0 and connects each q_i to the next q_{i+1}. Clearly, each p_n lies in $I^-(\gamma)$ and $I^-(\gamma) \subseteq W$. Using the openness of W and the denseness of the sequence $\{p_n\}$, it follows that $W = I^-(\gamma)$ as required. □

In space–times which are not strongly causal, there may exist future directed and future inextendible timelike curves γ such that $I^-(\gamma)$ is the chronological past of some point [i.e., $I^-(\gamma)$ is *not* a TIP]. Consider, for example, the cylinder $\mathbb{R}^1 \times S^1$ with the flat metric $ds^2 = dt\,d\theta$ and the usual time orientation with the future corresponding to increasing t. The lower half of the cylinder $W = \{(t, \theta) : t < 0\}$ is an IP which can be represented as $I^-(\gamma)$ for a future directed and future inextendible timelike curve γ. However, W is not a TIP since W can be represented as the chronological past of any point on the circle $t = 0$. By restricting our attention to strongly causal space–times, the IP's which are not TIP's are in one-to-one correspondence with the points of M. The dual statement holds for IF's which are not TIF's.

We now define $\hat{M}$ (respectively, $\check{M}$) to be the collection of all IP's (respectively, IF's). Furthermore, let $M^\sharp = \hat{M} \cup \check{M} / \sim$, where for each $p \in M$ the element $I^-(p)$ of $\hat{M}$ is identified with the element $I^+(p)$ of $\check{M}$. The map $I^+ : M \to M^\sharp$ given by $p \to I^+(p)$ then identifies M with a subset of $M^\sharp$. Using this identification, the set $M^\sharp$ corresponds to M together with all TIP's and TIF's.

In order to define a topology on $M^\sharp$, first define for any $A \in \check{M}$ the sets A^{int} and A^{ext} by

$$A^{\text{int}} = \{V \in \hat{M} : V \cap A \neq \emptyset\}$$

and

$$A^{\text{ext}} = \{V \in \hat{M} : V = I^-(W) \text{ implies } I^+(W) \not\subseteq A\}.$$

Similar definitions of B^{int} and B^{ext} are made for any $B \in \hat{M}$. A subbasis for a topology on $M^\sharp$ is then given by all sets of the form A^{int}, A^{ext}, B^{int}, and B^{ext}. The sets A^{int} and B^{int} are analogues of sets of the form $I^+(p)$ and $I^-(p)$, respectively. The sets A^{ext} and B^{ext} are analogues of $M - \overline{I^+(p)}$ and $M - \overline{I^-(p)}$, respectively.

Now in Geroch, Kronheimer, and Penrose (1972), it is proposed to obtain a Hausdorff space $M^* = M \cup \partial_c M$ from $M^\sharp$, with the topology given as above, by identifying the smallest number of points of $M^\sharp$ necessary to obtain a Hausdorff space M^*. Equivalently, it is proposed that $M^* = M \cup \partial_c M$ should be taken to be the quotient $M^\sharp / R_h$, where R_h is the intersection of all equivalence relations R on $M^\sharp$ such that $M^\sharp / R$ is Hausdorff.

Unfortunately, Szabados (1988) and Rube (1988, 1990) independently observed that the assumption of strong causality for the given space–time (M, g) was not sufficient to ensure that the minimal equivalence relation R_h exists, and hence the Geroch–Penrose–Kronheimer choice of topology may *not* be used to induce a Hausdorff causal completion M^* for a general strongly causal space–time. Szabados points out two difficulties that may arise for general strongly causal space–times if the topology given via $\{A^{\text{int}}, A^{\text{ext}}, B^{\text{int}}, B^{\text{ext}}\}$ is chosen. First, inner points corresponding to the embedding of M into $M^\sharp$ via the map I^+ and pre-boundary points are in general only T_1-separated, but not T_2-separated, with this choice of topology. Second, an example is given to show, for a nonspacelike curve γ with endpoint q in M, that the induced curve $I^+ \circ \gamma$ does not necessarily have a *unique* endpoint in M. Both Rube (1988) and Szabados (1988) suggest stronger causality conditions on the underlying space–time (M, g) which will ensure that the equivalence relation R_h proposed as above will exist, and hence this construction will produce a Hausdorff causal boundary $\partial_c M$ for M identified with its image in M^* under the mapping I^+. The simplest such condition to impose is that (M, g) be stably causal.

6.5 Local Extensions

In this section, extendibility and inextendibility of Lorentzian manifolds are defined. Also, two types of local extendibility are discussed. Most of the results of this section hold for Lorentzian manifolds which are not time orientable as well as for space–times.

Definition 6.15. (*Extension*) An *extension* of a Lorentzian manifold (M, g) is a Lorentzian manifold (M', g') together with an isometry $f : M \to M'$ which maps M onto a proper open subset of M'. An *analytic extension* of (M, g) is an extension $f : (M, g) \to (M', g')$ such that both Lorentzian manifolds are analytic and the map $f : M \to M'$ is analytic. If (M, g) has no extensions, it is said to be *inextendible*.

Suppose that the Lorentzian manifold (M, g) has an extension $f : (M, g) \to (M', g')$. Since M' is connected and $f(M)$ is assumed to be open in M', it follows that

$$\partial(f(M)) = \overline{f(M)} - f(M) \neq \emptyset,$$

where $\overline{f(M)}$ denotes the closure of $f(M)$ in M'. Because $\partial(f(M)) \neq \emptyset$ and the isometry $f : M \to M'$ maps geodesics in M into geodesics in M' lying in $f(M)$, it is easily seen that (M, g) cannot be timelike, null, or spacelike geodesically complete. Recalling that b-completeness and b.a. completeness both imply timelike geodesic completeness (cf. Section 6.2), we thus have the following criteria for Lorentzian manifolds to be inextendible.

Proposition 6.16. *A Lorentzian manifold (M, g) is inextendible if it is complete in any of the following ways:*

(1) *b-complete;*
(2) *b.a. complete;*
(3) *timelike geodesically complete;*
(4) *null geodesically complete; or*
(5) *spacelike geodesically complete.*

We now define two types of local extensions [cf. Clarke (1973, p. 207), Beem (1980), Hawking and Ellis (1973, p. 59)].

Definition 6.17. (*Local Extension*) Let (M, g) be a Lorentzian manifold.
(1) Suppose $\gamma : [0, a) \to M$ is a b-incomplete curve which is not extendible to $t = a$ in M. A *local b-boundary extension about* γ is an open neighborhood $U \subseteq M$ of γ and an extension (U', g') of $(U, g|_U)$ such that the image of γ in U' is C^0 extendible beyond $t = a$.
(2) A *local extension* of (M, g) is a connected open subset U of M having noncompact closure in M and an extension (U', g') of $(U, g|_U)$ such that the image of U has compact closure in U'.

Remark 6.18. This definition of local extension differs from the corresponding definition of local extension in Hawking and Ellis (1973, p. 59) in that U is required to be connected in Definition 6.17–(2) but not in Hawking and Ellis (cf. Figures 6.5 and 6.6).

We now investigate the relationships between these two types of local extensions. An arbitrary space–time may contain a b-incomplete curve $\gamma : [0, a) \to M$ which is not extendible to $t = a$, yet $\gamma[0, a)$ has compact closure in M. However, Schmidt has shown that such space–times contain compactly imprisoned inextendible null geodesics [cf. Schmidt (1973), Hawking and Ellis (1973, p. 280)]. On the other hand, if (M, g) contains no imprisoned nonspacelike curves, and (M, g) has a local b-boundary extension about γ, we now show this same extension yields a local extension.

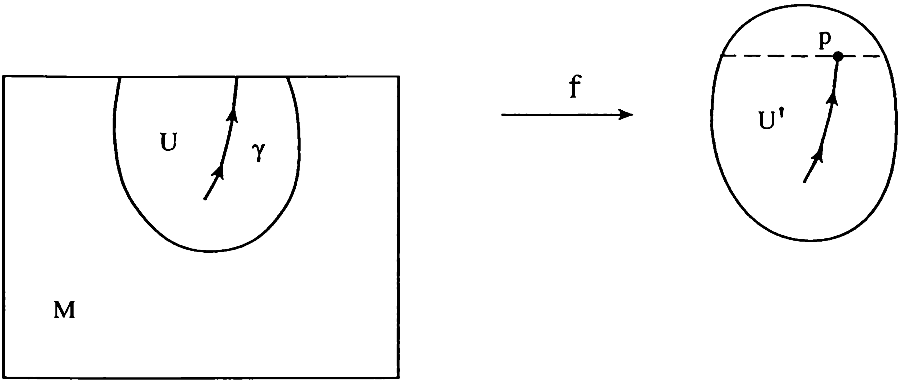


FIGURE 6.5. Let $\gamma : [0, a) \to M$ be a b-incomplete curve which is not extendible to $t = a$ in the space–time (M, g). Assume that there is an isometry $f : (U, g|_U) \to (U', g')$ which takes γ to a curve $f \circ \gamma$ having an endpoint p in U'. Then $f \circ \gamma$ may be continuously extended beyond $t = a$. Thus (M, g) has a local b-boundary extension about γ.

Lemma 6.19. *If (M, g) is a space–time with no imprisoned nonspacelike curves and if (M, g) has a local b-boundary extension about γ, then (M, g) has a local extension.*

Proof. Suppose that $f : (U, g|_U) \to (U', g')$ is a local b-boundary extension about γ. Then $f \circ \gamma : [0, a) \to U'$ is extendible, and $f \circ \gamma(t)$ converges to some $\overline{p} \in U'$ as $t \to a$. Let W' be an open neighborhood of $\overline{p}$ in U' with compact closure in U'. Choose $t_0 \in [0, a)$ such that $f \circ \gamma(t) \in W'$ for all $t_0 \leq t < a$. Set $V_1 = f^{-1}(W')$, and let V be the component of V_1 in U which contains the noncompact set $\gamma \,|\, [t_0, a)$. Since U is open in M, the set V is a connected open set in M with noncompact closure in M. Also $f(V)$ has compact closure in U' since $f(V) \subseteq W'$. Thus $f|_V : (V, g|_V) \to (U', g')$ is a local extension of (M, g). $\square$

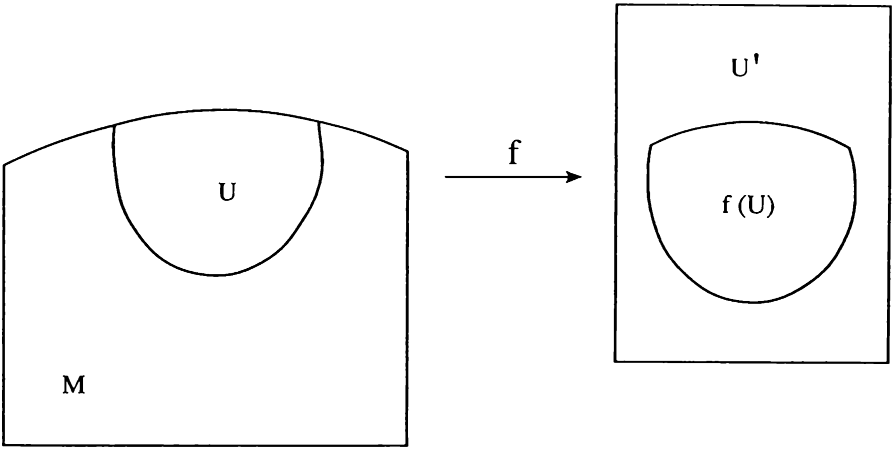


FIGURE 6.6. Let U be a connected open subset of M having non-compact closure in M. A local extension is defined to be an isometry $f : (U, g|_U) \to (U', g')$ such that $f(U)$ has compact closure in U'. Minkowski space–time shows that even real analytic b-complete space–times may admit analytic local extensions (cf. Example 6.21). Thus a space–time may admit local extensions but not admit local b-boundary extensions.

We now show that both types of local inextendibility imply global inextendibility.

Proposition 6.20. *If the Lorentzian manifold (M, g) has no local extensions of either of the two types in Definition 6.17, then (M, g) is inextendible.*

Proof. Suppose (M, g) has an extension $F : (M, g) \to (M', g')$. Let $\overline{p} \in \partial(F(M))$, and choose a geodesic $\sigma : [0, 1] \to (M', g')$ with $\sigma(0) \in F(M)$ and $\sigma(1) = \overline{p}$. Since $F(M)$ is open in M' and $\overline{p} \notin F(M)$, there exists some $t_0 \in (0, 1]$ such that $\sigma(t) \in F(M)$ for all $0 \leq t < t_0$ but $\sigma(t_0) \notin F(M)$. Then the curve $\gamma = F^{-1} \circ \sigma\big|_{[0,t_0)} : [0, t_0) \to M$ is b-incomplete, inextendible to

$t = t_0$ in M, and has noncompact closure in M. Taking $U = M$, $U' = M'$, and $f = F$ in (1) of Definition 6.17, it follows that (M, g) has a local b-boundary extension about γ. Taking W to be any open subset about $\overline{p}$ with compact closure in M', $U' = M'$, and U to be the component of $F^{-1}(W)$ containing γ, we obtain a local extension $F|_U : (U, g|_U) \to (M', g')$. □

The next example shows that Minkowski space–time has local extensions. Since Minkowski space–time is b-complete, this example shows that even though b-completeness is an obstruction to global extensions, it is *not* an obstruction to local extensions (cf. Proposition 6.16). This example is unusual in that it does *not* correspond to a local extension of M over a point of either the b-boundary $\partial_b M$ or the causal boundary $\partial_c M$. It is a local extension of a set which extends to i^0 (cf. Figure 5.4).

Example 6.21. Let $(M = \mathbb{R}^n, g)$ be n-dimensional Minkowski space–time and let $M' = \mathbb{R} \times T^{n-1}$, where $T^{n-1} = \{(\theta_2, \theta_3, \ldots, \theta_n) : 0 \leq \theta_i \leq 1\}$ is the $(n-1)$-dimensional torus (using the usual identifications). We may define a Lorentzian metric g' for M' by $g' = (ds')^2 = -dt^2 + d\theta_2{}^2 + \cdots + d\theta_n{}^2$. Then (M, g) is the universal Lorentzian covering space of (M', g') with covering map $f : (M, g) \to (M', g')$ given by

$$f(x_1, \ldots, x_n) = (x_1, x_2(\text{mod } 1), x_3(\text{mod } 1), \ldots, x_n(\text{mod } 1)).$$

Fix $\beta > 0$ and consider the curve $\gamma : [1, \infty) \to M$ given by

$$\gamma(s) = (s^{-\beta}, s, 0, \ldots, 0).$$

Then $f \circ \gamma : [1, \infty) \to M'$ is a spiral which is asymptotic to the circle $t = \theta_3 = \cdots = \theta_n = 0$ in M'. Let U be an open tubular neighborhood about γ in M such that U is contained in some open set $\{(x_1, \ldots, x_n) \in \mathbb{R}^n : 0 < x_1 < \alpha\}$ for some fixed $\alpha > 1$ and such that $f|_U : U \to M'$ is a homeomorphism onto its image (cf. Figure 6.7). Intuitively, the set U must be chosen to be thinner as $s \to \infty$ in order to satisfy the requirement $x_1 > 0$ for $(x_1, \ldots, x_n) \in U$. While U does not have compact closure in Minkowski space–time, $f(U)$ does have compact closure in M' since $f(U)$ is contained in the compact set $[0, \alpha] \times T^{n-1}$.

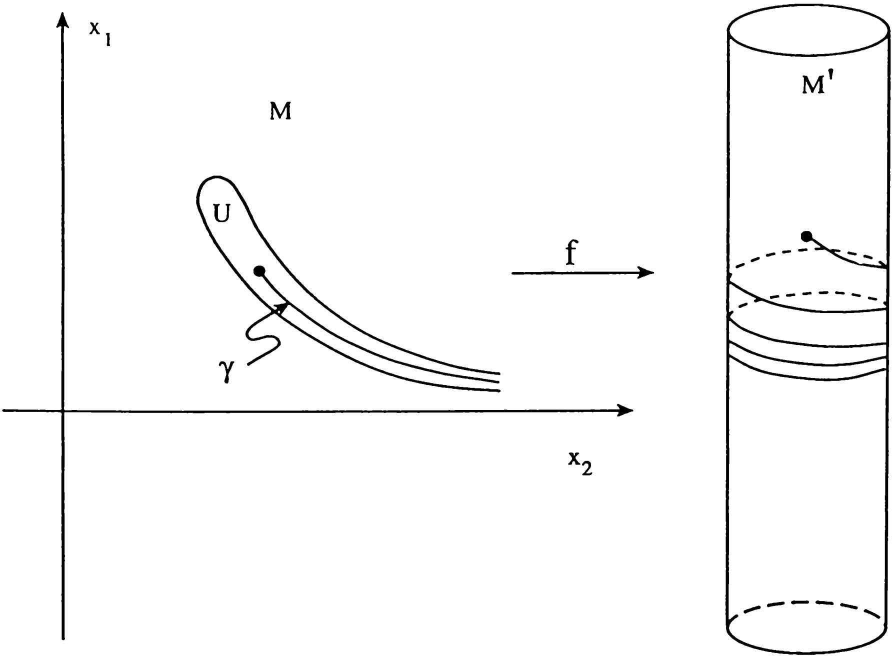


FIGURE 6.7. Minkowski space–time has analytic local extensions. Let $M = \mathbb{R}^n$ be given the usual Minkowskian metric g, and let T^{n-1} be the $(n-1)$-dimensional torus with the usual positive definite flat metric h. Let $M' = \mathbb{R} \times T^{n-1}$ be given the Lorentzian product metric $g' = -dt^2 \oplus h$. Then (M, g) is the universal covering space of (M', g'), and the quotient map $f : M \to M'$ is locally isometric. Choose U to be an open set in M about $\gamma(s) = (s^{-\beta}, s, \ldots, 0)$, $\gamma : [1, \infty) \to M$, such that $f \mid U$ is one-to-one and $f(U)$ has compact closure in M'. Then $f|_U : (U, g_U) \to (M', g')$ is an analytic local extension of Minkowski space, but this extension is *not* across points of $\partial_c M$ and *not* across points of $\partial_b M$.

Thus $f : (U, g|_U) \to (M', g')$ is an analytic local extension of Minkowski space–time. Notice that if $\gamma_1 : [0, a) \to U$ is any curve with noncompact closure in M, then $f \circ \gamma_1$ cannot be extended to $t = a$ in M'.

6.6 Singularities

Let ∂M denote an ideal boundary of M (i.e., ∂M represents either $\partial_b M$ or $\partial_c M$). A point $q \in \partial M$ is said to be a *regular boundary point* of M if there exists a global extension (M', g') of (M, g) such that q may be naturally identified with a point of M'. A regular boundary point may thus be regarded as being a removable singularity of M.

Let $\gamma : [0, a) \to M$ be an inextendible curve such that $\gamma(a)$ corresponds to an ideal point of M. The curve γ is said to define a *curvature singularity* [cf. Ellis and Schmidt (1977, p. 916)] if some component of $R_{abcd;e_1,\ldots,e_k}$ is *not* C^0 on $[0, a]$ when measured in a parallelly propagated orthonormal basis along γ. A curvature singularity is an obstruction to a local b-boundary extension about γ because if there is a local b-boundary extension about γ, then the curvature tensor and all of its derivatives measured in a parallelly propagated orthonormal basis must be continuous and hence converge to well-defined limits as $t \to a^-$. A related but somewhat different notion is that of *strong curvature singularity* which may be defined using expansion $\overline{\theta}$ along null geodesics. The notion of strong curvature singularity has been useful in connection with cosmic censorship [cf. Królak (1992), Królak and Rudnicki (1993)].

A b-boundary point $q \in \partial_b M$ which is neither a regular boundary point nor a curvature singularity is called a *quasi-regular singularity*. Clarke (1973, p. 208) has proven that if $\gamma : [0, a) \to M$ is an inextendible b-complete curve which corresponds to a quasi-regular singularity, then there is a local b-boundary extension about γ. This shows that curvature singularities are the only real obstructions to local b-boundary extensions.

In general, it can be quite difficult to decide if a given space–time has local extensions of some type. However, for analytical local b-boundary extensions of analytic space–times, the situation is somewhat simpler (cf. Theorem 6.23).

For the proof of Theorem 6.23, it is useful to prove the following proposition about real analytic space–times and local isometries. Recall that a local isometry $F : M \to M'$ is a map such that for each $p \in M$, there exists an open neighborhood $U(p)$ of p on which F is an isometry. Thus local isometries are local diffeomorphisms but need not be globally one-to-one.

Proposition 6.22. *Let (M, g) and (M_1, g_1) be real analytic space–times of the same dimension and suppose that $F : M \to M_1$ is a real analytic map. If M contains an open set U such that $F|_U : U \to M_1$ is an isometry, then F is a local isometry.*

Proof. Let $W = \{m \in M : F_*v \neq 0 \text{ for all } v \neq 0 \text{ in } T_mM\}$, which is an open subset of M by the inverse function theorem. Since $F|_U$ is an isometry, U is contained in W. Fix any $p \in U$, and let V be the path connected component of W containing p. We will establish the proposition by showing first that $F|_V$ is a local isometry and second that $V = M$.

Let q be any point of V. Choose a curve $\gamma : [0,1] \to V$ with $\gamma(0) = p$ and $\gamma(1) = q$. By the usual compactness arguments, we may cover $\gamma[0,1]$ with a finite chain of coordinate charts $(U_1, \phi_1), (U_2, \phi_2), \ldots, (U_k, \phi_k)$ such that each U_i is simply connected, $F|_{U_i} : U_i \to M_1$ is an analytic diffeomorphism, $p \in U_1 \subseteq U \cap V$, $q \in U_k$, and $U_i \cap U_{i+1} \neq \emptyset$ for each i with $1 \leq i \leq k-1$. Since $U_1 \subseteq U \cap V$, we have $g = (F|_{U_1})^* g_1$ on U_1. Thus $g = (F|_{U_1})^* g_1$ on $U_1 \cap U_2$. Since $U_1 \cap U_2$ is an open subset of U_2 and F is a real analytic diffeomorphism of U_2 onto its image, it follows that $g = (F|_{U_2})^* g_1$ on U_2. Continuing inductively, we obtain $g = (F|_{U_k})^* g_1$ on U_k, whence F is an isometry in the open neighborhood U_k of q. Thus $F|_V : V \to M_1$ is a local isometry.

It remains to show that $V = M$. Suppose $V \neq M$. Choose any point $r_1 \in M - V$. Let $\gamma_1 : [0,1] \to M$ be a smooth curve with $\gamma(0) = p$ and $\gamma(1) = r_1$. There is a smallest $t_0 \in [0,1]$ such that $r = \gamma(t_0) \in M - V$. Then F restricted to the neighborhood V of $\gamma_1[0, t_0)$ is a local isometry. It suffices to show that $r \in V$ to obtain the desired contradiction. Since $r \in M - V$, there exists a tangent vector $x \neq 0$ in T_rM with $F_*x = 0$. Let X be the unique parallel field along γ with $X(t_0) = x$. Then $F_* \circ X$ is a parallel field along $F \circ \gamma_1 : [0, t_0) \to M_1$ since F is a local isometry in a neighborhood of

$\gamma_1 : [0, t_0) \to M$. But since F is smooth, $F_* x = F_* X(t_0) = \lim_{t \to t_0^-} F_* X(t)$. Because $F_* \circ X$ is a parallel vector field for all t with $0 \leq t \leq t_0$, it follows that $\lim_{t \to t_0^-} F_* X(t) \neq 0$. Hence $F_* x \neq 0$. Thus F is nonsingular at the point r, whence $r \in V$ in contradiction. □

We are now ready to turn to the proof of Theorem 6.23 on local b-boundary extensions of real analytic space–times.

Theorem 6.23. *Suppose (M, g) is an analytic space–time with no imprisoned nonspacelike curves, which has an analytic local b-boundary extension about $\gamma : [0, a) \to M$. Then there are timelike, null, and spacelike geodesics of finite affine parameter which are inextendible in one direction and which do not correspond to curvature singularities.*

The proof of Theorem 6.23 will involve two lemmas.

Lemma 6.24. *Suppose (M, g) is an analytic space–time with no imprisoned nonspacelike curves, which has an analytic local b-boundary extension about $\gamma : [0, a) \to M$. Then (M, g) has an incomplete geodesic.*

Proof. Let $f : (U, g|_U) \to (U', g')$ be an analytic extension about γ. We may assume U contains the image of γ. Also, $f \circ \gamma$ is extendible in U'. Thus $f \circ \gamma(t) \to p \in U'$ as $t \to a^-$. Let W' be a neighborhood of p such that W' is a convex normal neighborhood of each of its points. Then $\exp_x^{-1} : W' \to T_x U'$ is a diffeomorphism for each fixed $x \in W'$. Assume t_0 is chosen with $f \circ \gamma(t) \in W'$ for all $t_0 \leq t < a$. Set $q = \gamma(t_0)$ and $r = f(q)$. Then $H = \exp_q \circ f_{*_q}^{-1} \circ \exp_r^{-1} : W' \to M$ is analytic and is at least defined near r. The map H takes geodesics starting at r to geodesics starting at q, and H preserves lengths along these geodesics. In fact, H agrees with f^{-1} near r. The map H need not be one-to-one since $\exp_q$ is not necessarily one-to-one. Because the domain of $\exp_q$ is a union of line segments starting at the origin of $T_q M$, the domain V' of H must be some subset of W' which is a union of geodesic segments starting at r. Hence the set V' fails to be all of W' only when $\exp_q : T_q M \to M$ is defined on a proper subset of $T_q M$ which does not include all of the image $f_{*_q}^{-1} \circ \exp_r^{-1}(W')$. Thus if we show $V' \neq W'$, there is some incomplete, inextendible geodesic starting at q. But the analytic maps

H and f^{-1} must agree on the component of $f(U) \cap V'$ which contains r. This implies $V' \neq W'$. Otherwise, H and f^{-1} would agree on $f(U) \cap W'$ and hence on a neighborhood of $f \circ \gamma[t_0, a)$. This yields $H \circ f \circ \gamma = \gamma$ for $t_0 \leq t < a$ and implies γ is extendible in M across the point $H(p)$, in contradiction. □

We will continue with the same notation in the next lemma.

Lemma 6.25. *The map $H : V' \to M$ is a local isometry.*

Proof. The space–time (M, g) is analytic, (U', g') is analytic, and H is analytic. Furthermore, H agrees with the isometry f^{-1} near r, and H is defined on an arcwise connected set V'. Thus Proposition 6.22 implies H is a local isometry. □

We are now ready to complete the

Proof of Theorem 6.23. There are three cases to consider corresponding to incomplete timelike, null, and spacelike geodesics. We only give the proof for the timelike case. Let U, U', f, etc., be as in Lemmas 6.24 and 6.25. Assume without loss of generality that there is some point $x \in W'$ such that in a chronological ordering on W', we have $x \ll p$ and $x \ll f \circ \gamma(t)$ for all $t_0 \leq t < a$ (cf. Figure 6.8).

If $x \notin V'$, let α be the geodesic segment in W' from r to x. Then H takes $\alpha \cap V'$ to an inextendible, incomplete, timelike geodesic starting at q. Lemma 6.25 implies that this geodesic does not correspond to a curvature singularity.

If $x \in V'$, let $y = H(x)$ and define $H' = \exp_y \circ H_{*_x} \circ \exp_x^{-1} : W' \to M$. The map H' is defined on some subset V'' of W'. It is a local isometry for the same reasons that H is a local isometry, and H' agrees with both H and f^{-1} near r. The set V'' cannot contain all of $f \circ \gamma$ for γ on $t_0 \leq t < a$ since this would yield an endpoint $H'(p)$ of γ in M. Using $x \ll f \circ \gamma(t)$ for $t_0 \leq t < a$, we conclude that there is an inextendible incomplete timelike geodesic starting at y in M which does not correspond to a curvature singularity. □

Remark 6.26. There are examples of C^∞ space–times which are both geodesically complete and locally b-extendible [cf. Beem (1976c, p. 506)].

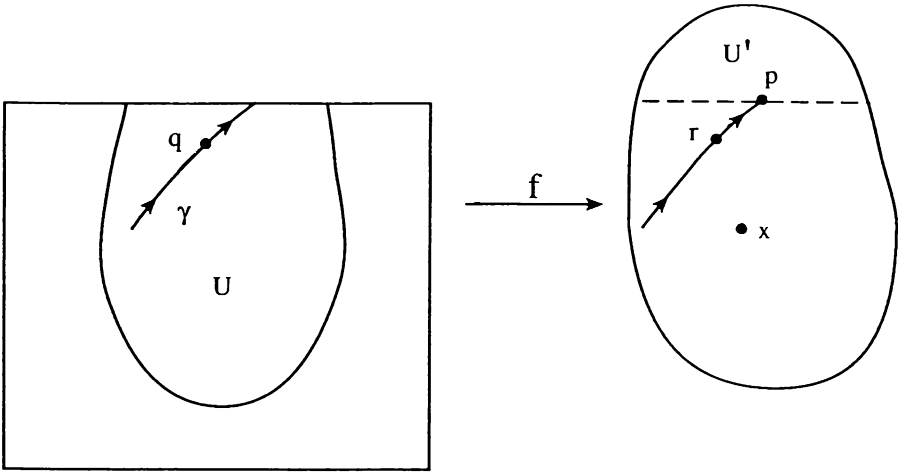


FIGURE 6.8. In the proof of Theorem 6.23, the map $f : (U, g|_U) \to (U', g')$ is an isometry which is an analytic local b-boundary extension about γ. The point $p \in U'$ is the endpoint of $f \circ \gamma$ in U'. In this figure, $f(q) = r$, and all points of $f \circ \gamma$ between r and p are in the chronological future of x.

Hence the analyticity in the hypothesis of Theorem 6.23 cannot be replaced by a C^∞ assumption.

Corollary 6.27. *Let (M, g) be an analytic space–time with no imprisoned nonspacelike curves such that each timelike geodesic $\gamma : [0, a) \to M$ which is inextendible to $t = a$ is either complete (i.e., $a = \infty$) in the indicated direction or else corresponds to a curvature singularity. Then (M, g) has no analytic local b-boundary extensions.*

Extensive work has been done on classifying singularities by various different methodologies. Indeed, the monograph by Clarke (1993) may be consulted to obtain much more detailed information than that which is provided in this chapter. Our treatment here will be limited to a discussion of the "abstract boundary" $\partial_a(M)$ and its application to the classification of singularities as given in Scott and Szekeres (1994) and Fama and Scott (1994). (In Clarke (1993), a somewhat different "A-boundary" obtained by taking the closures of an atlas of normal coordinate charts is discussed.) To form the a-boundary, equivalence classes of boundary points of a given smooth manifold under all possible open embeddings are considered. The a-boundary thus has no dependence on any choice of affine connection or semi-Riemannian metric. When the manifold is further endowed with an extra structure, like an affine connection or a semi-Riemannian metric, then abstract boundary points may be classified as regular boundary points, points at infinity, unapproachable points, or singularities, with the classification having some dependence upon a particular curve family $\mathcal{C}$ selected, which is associated to the added structure for M. The abstract boundary also has the feature that if a closed region is removed from a singularity-free semi-Riemannian manifold, then the resulting semi-Riemannian manifold is still singularity-free since only regular boundary points are introduced by the excision of the closed set.

We will begin by discussing the construction of the a-boundary for a given smooth manifold M of dimension n. Since no metric is yet involved, we will adopt the terminology of Scott and Szekeres (1994) and define an *enveloped manifold* (M, M_1, f_1) to be a pair of (connected) smooth manifolds M, M_1 of the same dimension together with a smooth embedding $f_1 : M \to M_1$. Since both manifolds have the same dimension, $f_1(M)$ is an open submanifold of M_1. (A very special type of envelopment would then be provided by a global extension of a given space–time [cf. Definition 6.15].)

Definition 6.28. (*Boundary Point, Boundary Set*) A *boundary point* p of an enveloped manifold (M, M_1, f_1) is a point $p \in M_1 - f_1(M)$ such that every open neighborhood U of p in M_1 has non-empty intersection with $f_1(M)$, i.e., p belongs to the topological boundary $\partial(f_1(M)) = \overline{f_1(M)} - f_1(M)$. A *boundary set* B contained in $M_1 - f_1(M)$ is any set of boundary points for the enveloped manifold (M, M_1, f_1). We will use the notation (M, M_1, f_1, B) for an enveloped manifold with distinguished boundary set B.

The set of enveloped manifolds and resulting boundary sets is much too large an object to use to form a useful boundary. Thus the following relation is introduced between enveloped manifolds and boundary sets.

Definition 6.29. (*Covering Boundary Sets*) Let (M, M_1, f_1, B_1) and (M, M_2, f_2, B_2) be two envelopments for M with boundary sets B_1 and B_2, respectively. Then the boundary set B_1 is said to *cover* the boundary set B_2 if for every open neighborhood U_1 of B_1 in M_1, there exists an open neighborhood U_2 of B_2 in M_2 such that

$$f_1 \circ f_2^{-1}\left(U_2 \cap f_2(M)\right) \subseteq U_1. \tag{6.2}$$

This requirement may be conveniently checked in terms of sequences as follows.

Proposition 6.30. *The boundary set B_1 covers the boundary set B_2 iff for every sequence $\{p_k\}$ in M such that the sequence $\{f_2(p_k)\}$ has a limit point in B_2, the sequence $\{f_1(p_k)\}$ is required to have a limit point in B_1.*

Since the requirement (6.2) is to some degree only a topological restriction, examples may be given of different envelopments of $M = \mathbb{R}^n - \{(0)\}$ for which one envelopment produces the boundary set $B_1 = \{(0)\}$ consisting of precisely one point, but a second envelopment produces the boundary set $B_2 = S^{n-1}$, and even more remarkably, B_1 covers B_2 and B_2 covers B_1. Thus a single point which is the boundary set of a given envelopment may cover *infinite* sets of points which are contained in a boundary set for a second envelopment.

Definition 6.31. (*Equivalent Boundary Sets*) Let (M, M_1, f_1, B_1) and (M, M_2, f_2, B_2) be two enveloped manifolds with boundary sets B_1 and B_2, respectively. Then B_1 is said to be *equivalent* to B_2 iff B_1 covers B_2 and B_2 covers B_1.

Denote this equivalence relation on the class of boundary sets for M by $B_1 \sim B_2$. An *abstract boundary set* $[B]$ is then an equivalence class of boundary sets. In Fama and Scott (1994), topological properties of boundary set equivalence are studied. It is noted that if a boundary set B of a given envelopment is compact, then all boundary sets B' of all other envelopments which are equivalent to B must also be compact. In particular, all boundary sets B equivalent to a single boundary point $[p]$ are compact. However, closedness, connectedness, and simple connectedness of boundary sets are unfortunately *not* invariant under boundary set equivalence. Fama and Scott (1994) give a detailed discussion of "topological neighborhood properties," including the connected neighborhood property and the simply connected neighborhood property, which are preserved by boundary set equivalence.

With all of these preliminaries settled, we may now define the *a-boundary* as in Scott and Szekeres (1994).

Definition 6.32. (*The a-boundary* $\partial_a(M)$) If p is a boundary point of an enveloped manifold (M, M_1, f_1), then the equivalence class $[p]$ of the boundary set $\{p\}$ under the equivalence relation $\sim$ of Definition 6.31 is called an *abstract boundary point of* M. The set $\partial_a(M)$ of all such abstract boundary points for all envelopments of the given manifold M is called the *abstract boundary* or the *a-boundary*.

More symbolically, one has that

$$\partial_a(M) = \{[p] : p \text{ is in } \overline{f_1(M)} - f_1(M) \text{ for some envelopment } (M, M_1, f_1)\}.$$

In the beginning of this section, the concept of regular boundary point was defined for space–times in terms of global extensions. A similar definition is given in Scott and Szekeres (1994) for general semi-Riemannian manifolds. Now let (M, g) be a semi-Riemannian manifold with a metric tensor g of class C^k. Let (M, g, M_1, f_1) be an envelopment of M. The induced metric tensor $\left(f_1^{-1}\right)^* g$ on $f_1(M)$ will be denoted by g when there is no risk of ambiguity.

Definition 6.33. (C^l*-Extension of* (M, g)) A C^l*-extension* $(1 \leq l \leq k)$ of a C^k semi-Riemannian manifold (M, g) is an envelopment of (M, g) by a C^l semi-Riemannian manifold (M_1, g_1), $f_1 : M \rightarrow M_1$, such that

$$g_1 \mid f_1(M) = g.$$

This will be denoted by (M, g, M_1, g_1, f_1).

With this definition in hand, the concept of a C^l*-regular boundary point* may be formulated independent of any choice of curve family for (M, g).

Definition 6.34. (*C^l-Regular Boundary Point*) A boundary point p of an envelopment (M, g, M_1, g_1, f_1) is *C^l-regular for g* if there exists a C^l semi-Riemannian manifold (M_2, g_2) such that $f_1(M) \cup \{p\} \subseteq M_2 \subseteq M_1$ and (M, g, M_2, g_2, f_1) is a C^l-extension of (M, g).

Further analysis of the boundary points as "singular boundary points" or "points at infinity" is possible only if a certain family of curves $\mathcal{C}$ with the bounded parameter property has been specified, and indeed it is natural that this classification should depend on the choice of the particular curve family selected. Thus we assume that the smooth manifold M is furnished with a family $\mathcal{C}$ of parametrized curves satisfying the bounded parameter property. Here a parametrized curve will be taken to mean a C^1 map $\gamma : [a, b) \to M$ with $a < b \leq +\infty$ and with $\gamma'(t) \neq 0$ for all t in $[a, b)$.

Definition 6.35. (*Bounded Parameter Property*) A family $\mathcal{C}$ of parametrized curves in M is said to satisfy the *bounded parameter property* if the members of $\mathcal{C}$ satisfy the following:

(1) Through any point p of M passes at least one curve γ of the family $\mathcal{C}$;
(2) If γ is a curve in $\mathcal{C}$, then so is every subcurve of γ; and
(3) For any pair of curves γ_1 and γ_2 in $\mathcal{C}$ which are obtained from each other by a monotone increasing C^1 change of parameter, either the parameter on both curves is bounded, or it is unbounded on both curves.

Examples of suitable families of curves $\mathcal{C}$ in our context include: (1) the family of all geodesics with affine parameter in a manifold M with affine connection; (2) all C^1 curves parametrized by "generalized affine parameter" (cf. Definition 6.8) in an affine manifold; and (3) all future timelike geodesics parametrized by arc length in a space–time.

As an alternative to Proposition 6.30, boundary sets may be studied using parametrized curves.

Definition 6.36. (*Limit Point, Endpoint of Curve*) Let $\gamma : [a, b) \to M$ be a parametrized curve. Then

(1) A point p in M is a *limit point* of the curve $\gamma : [a, b) \to M$ if there

exists an increasing sequence $t_i \to b$ such that $\gamma(t_i) \to p$.

(2) A point p in M is an *endpoint* of the curve γ if $\gamma(t) \to p$ as $t \to b^-$. (For Hausdorff manifolds, curve endpoints are unique.)

Definition 6.37. (*Approaching a Boundary Set*) Let $\gamma : [a, b) \to M$ be a parametrized curve and (M, M_1, f_1, B) a boundary set in an envelopment of M.

(1) The parametrized curve $\gamma : [a, b) \to M$ *approaches the boundary set* B if the curve $f_1 \circ \gamma$ has a limit point lying in B.

(2) The parametrized curve $\gamma : [a, b) \to M$ *has its endpoint in* B if the curve $f \circ \gamma : [a, b) \to M_1$ has its endpoint in B.

The analogue of Proposition 6.30 above in this setting is then

Proposition 6.38. *If a boundary set B_1 covers a boundary set B_2, then every curve $\gamma : [a, b) \to M$ which approaches B_2 also approaches B_1.*

Now we restrict our attention to a semi-Riemannian manifold (M, g) for which a family $\mathcal{C}$ of curves for M with the bounded parameter property has been selected, and we use the notation $(M, g, \mathcal{C})$ to denote this selection.

Definition 6.39. (*Approachable Boundary Point, $\mathcal{C}$-Boundary Point*) If (M, M_1, f_1) is an envelopment of $(M, g, \mathcal{C})$, then a boundary point p of this envelopment is *approachable*, or a *$\mathcal{C}$-boundary point*, if p is a limit point of some curve $\gamma : [a, b) \to M$ of the family $\mathcal{C}$. Boundary points which are *not* $\mathcal{C}$-boundary points will be called *unapproachable*.

As a result of Proposition 6.38, this definition passes to the a-boundary:

Definition 6.40. (*Approachable a-Boundary Point, Abstract $\mathcal{C}$-Boundary Point*) An abstract boundary point $[p]$ is an *abstract $\mathcal{C}$-boundary point*, or *approachable*, if p is a $\mathcal{C}$-boundary point. Similarly, an abstract boundary point $[p]$ is *unapproachable* if p is not a $\mathcal{C}$-boundary point.

With the family $\mathcal{C}$ specified, the points at infinity for $(M, g, \mathcal{C})$ may now be detected by determining whether they may be reached along any curve in $\mathcal{C}$ at a *finite* value of the given parameter.

Definition 6.41. (*C^l-Point at Infinity for $(M, g, \mathcal{C})$*) Given $(M, g, \mathcal{C})$, a boundary point p of an envelopment $(M, g, \mathcal{C}, M_1, f_1)$ is said to be a *C^l-point at infinity for $\mathcal{C}$* if

(1) p is not a C^l-regular boundary point (recall Definition 6.34),
(2) p is a $\mathcal{C}$-boundary point, and
(3) no curve in $\mathcal{C}$ approaches p with bounded parameter.

Since the family $\mathcal{C}$ is assumed to satisfy the bounded parameter property, the concept of a point at infinity is independent of the choice of parametrization for the curves from $\mathcal{C}$ which approach p. Condition (3) says more explicitly that for no interval $[a, b)$ with b finite is there a curve $\gamma : [a, b) \to (M, g)$ in the family $\mathcal{C}$ and an increasing sequence of real numbers $t_i \to b$ such that

$$f_1(\gamma(t_i)) \to p \text{ as } t_i \to b.$$

A tricky aspect of a point p at infinity for $(M, g, \mathcal{C})$ is that it is possible that p is covered by a boundary set B of another embedding consisting only of regular or unapproachable boundary points. In this case, Scott and Szekeres (1994, p. 238) term the point p at infinity for $(M, g, \mathcal{C})$ a *removable* point at infinity. If no such covering by a boundary set of another embedding exists, then p is termed an *essential point at infinity*. It may be shown that the concept of being an essential point at infinity passes to the abstract boundary, i.e., these points cannot be transformed away by a change of coordinates. Thus an abstract boundary point $[p]$ is termed an *abstract point at infinity* if it has a representative in some envelopment which is an essential point at infinity for that envelopment. Essential points at infinity may cover regular boundary points of another embedding. Hence, the following final dichotomy is used for essential points at infinity: an essential boundary point at infinity which covers a regular boundary point is termed a *mixed point at infinity.* Otherwise, p is termed a *pure point at infinity*.

So far the two categories of regular points and points at infinity have been defined. The final category to be treated is that of singular boundary points. Here a more subtle viewpoint is taken than that of Definition 6.3 above in which "geodesic singularity" is defined. According to this previous definition, if one takes, for example, two points p and q in Minkowski space with $p \ll q$ and puts $M = I^+(p) \cap I^+(q)$ with the induced metric from the inclusion of M in Minkowski space, then M is a singular space–time. Indeed, every geodesic of (M, g) is incomplete. Yet the given space–time has a global embedding in Minkowski space such that after this enlargement, every geodesic becomes complete. Thus this example should perhaps not really be regarded as a singular space–time, Definition 6.3 notwithstanding.

Definition 6.42. (*C^l-Singular Point*) A boundary point p of an envelopment $(M, g, \mathcal{C}, M_1, f_1)$ is said to be *C^l-singular* if

(1) p is not a C^l-regular boundary point,
(2) p is a $\mathcal{C}$-boundary point, and
(3) there exists a curve in the family $\mathcal{C}$ which approaches p with bounded parameter.

Thus a singular boundary point for the envelopment is a $\mathcal{C}$-boundary point which is not C^l-regular and is not a point at infinity. Again, in this case, a finer subclassification is made. If p is covered by a non-singular boundary set of a second envelopment, then p is termed a *removable singularity*. If not, then p is an *essential singularity*, and a further classification is made as follows: if p covers some regular boundary points or points at infinity of another embedding, then p is commonly called a *mixed* or *directional singularity*. If no such covering behavior is exhibited, then p is said to be a *pure singularity*. Then an abstract boundary point $[p]$ in $\partial_a(M)$ may be termed an *abstract singularity* if it has a representative which is an essential singularity.

We conclude this section with considerations akin to the more simple concepts of Section 6.2 and "geodesic singularity" of Definition 6.3. First, the notion of geodesic completeness for a space–time may be extended to that of $\mathcal{C}$-completeness for the manifold M with distinguished curve family $\mathcal{C}$.

Definition 6.43. (*$\mathcal{C}$-Completeness*) Given a manifold M with a curve family $\mathcal{C}$ satisfying the bounded parameter property, M is said to be *$\mathcal{C}$-complete* if every curve $\gamma : [a, b) \to M$ in $\mathcal{C}$ with bounded parameter, i.e., $b < +\infty$, has an endpoint in M.

Definition 6.44. (C^l*-Singularity*) A semi-Riemannian manifold (M, g) with a distinguished class $\mathcal{C}$ of curves satisfying the bounded parameter property has a C^l*-singularity* if there exists an envelopment of M having an essential C^l-singularity p, i.e., p is a C^l-singular point for some envelopment of $(M, g, \mathcal{C})$ which is *not* covered by a C^l-non-singular boundary set B of any other envelopment for $(M, g, \mathcal{C})$. Conversely, $(M, g, \mathcal{C})$ is said to be C^l*-singularity free* if it has no C^l-singularities, i.e., for every envelopment of M, the boundary points are either C^l-nonsingular (C^l-regular boundary points, C^l-points at infinity, or unapproachable boundary points), or C^l-removable singularities.

According to Scott and Szekeres (1994), any theory of singularities ought to pass over the following hurdle.

Theorem 6.45. *Every compact semi-Riemannian manifold* $(M, g, \mathcal{C})$*, with any family of curves* $\mathcal{C}$ *satisfying the bounded parameter property, is singularity free.*

Proof. A compact manifold M has no non-trivial envelopments (M, M_1, f_1), for M_1 is required to be connected yet would contain $f_1(M)$ as a compact open subset. Thus $f_1(M)$ would be both open and closed in M_1, whence $M_1 = f_1(M)$. Since M has no envelopments, $\partial_a(M)$ is empty and hence can contain no singular points. □

Theorem 6.46. *If* (M, g) *is a semi-Riemannian manifold which is* $\mathcal{C}$*-complete for a curve family* $\mathcal{C}$ *satisfying the bounded parameter property, then* $(M, g, \mathcal{C})$ *is singularity-free.*

Proof. Suppose that $(M, g, \mathcal{C})$ contains a singularity. Then there exists an envelopment $(M, g, \mathcal{C}, M_1, f_1)$ which has a $\mathcal{C}$-boundary point $p \in M_1 - f_1(M)$. Hence, there exists a curve $\gamma : [a, b) \to M$ in $\mathcal{C}$ which has p as a limit point and also $b < +\infty$. Now by the assumed $\mathcal{C}$-completeness, γ has an *endpoint* q in M. But endpoints are unique limit points of curves so that of necessity $p = f_1(q)$, which is impossible since $p \notin f_1(M)$. □

In Section 6.5, Proposition 6.16, it was noted that various types of completeness such as b-completeness, b.a. completeness, timelike geodesic completeness, null geodesic completeness, and spacelike geodesic completeness preclude global extendibility. In the present context, the analogous statement is the following.

Theorem 6.47. *If the semi-Riemannian manifold* (M, g) *is geodesically complete, then* (M, g) *has no regular boundary points.*

Complete discussions of all of these boundary points classifications together with extensive examples, including Taub–NUT space–time as simplified by Misner (1967) and the Curzon space–time, may be found in Scott and Szekeres (1994) and Fama and Scott (1994).

CHAPTER 7

STABILITY OF COMPLETENESS AND INCOMPLETENESS

In proving singularity theorems in general relativity, it is important to use hypotheses that hold not just for the given "background" Lorentzian metric g for M but in addition for all metrics g_1 for M sufficiently close to g. Not only does the imprecision of astronomical measurements mean that the Lorentzian metric of the universe cannot be determined exactly, but also cosmological assumptions like the spatial homogeneity of the universe hold only approximately. Nevertheless, if an incompleteness theorem can be obtained for the idealized model (M, g) using hypotheses valid for all metrics g_1 for M in an open neighborhood of g, then all space–times (M, g_1) with g_1 sufficiently close to g will also be incomplete. Hence if the model is believed to be sufficiently accurate, conclusions valid for the model are also valid for the actual universe.

Recall that $\text{Lor}(M)$ *denotes the space of all Lorentzian metrics for a given manifold* M and that $\text{Con}(M)$ *denotes the quotient space formed by identifying all pointwise globally conformal metrics* $g_1 = \Omega g_2$ for M, where $\Omega : M \to (0, \infty)$ is smooth. Let $\tau : \text{Lor}(M) \to \text{Con}(M)$ denote the natural projection map which assigns to each Lorentzian metric g for M the set $\tau(g) = \hat{g}$ of all Lorentzian metrics for M pointwise globally conformal to g. Given $\hat{g} \in \text{Con}(M)$, set $C(M, g) = \tau^{-1}(\hat{g}) \subseteq \text{Lor}(M)$. It is customary in general relativity to say that a curvature or causality condition for a space–time (M, g) is C^r *stable* in $\text{Lor}(M)$ [respectively, $\text{Con}(M)$], if the validity of the condition for (M, g) implies the validity of the condition for all g_1 in a C^r-open neighborhood of g in $\text{Lor}(M)$ [respectively, $\text{Con}(M)$]. More generally, *a stable condition for a set of metrics is one which holds on an open subset of such metrics.*

After the singularity theorems described in Chapter 8 of Hawking and Ellis (1973) were obtained, it was of interest to study the C^r stability of conditions such as the existence of closed trapped surfaces, positive nonspacelike Ricci curvature, and geodesic completeness, which played such a key role in these singularity theorems. Geroch (1970a) established the stability of global hyperbolicity in the interval topology on Con(M). Then Lerner (1973) made a thorough study of the stability in Lor(M) and Con(M) of causality and curvature conditions useful in general relativity. In particular, Lerner noted that the interval and quotient topologies for Con(M) coincide. Hence Geroch's stability result for global hyperbolicity holds for Con(M) in the quotient topology and thus automatically holds in Lor(M). In Section 7.1 we define the fine C^r topologies and the interval topology for Con(M). We then review stability properties of Lor(M) and Con(M) which were established by Geroch (1970a) and Lerner (1973). We give examples of Williams (1984) to show that both *geodesic completeness and geodesic incompleteness may fail to be stable.* These two properties are C^0 stable for definite spaces, but for all signatures (s, r) with $s \geq 1$ and $r \geq 1$ one may construct examples for which these properties are unstable.

In Section 7.2, using the "Euclidean norm"

$$\|\xi - \eta\|_2 = \left(\sum_{i=1}^{2n} [\,\widetilde{x}_i(\xi) - \widetilde{x}_i(\eta)]^2 \right)^{\frac{1}{2}}$$

induced on $(TM|_U, \widetilde{x})$ by a coordinate chart (U, x) for M and standard estimates from the theory of systems of ordinary differential equations in $\mathbb{R}^n$, we obtain estimates for the behavior of geodesics in (U, x) under C^1 metric perturbations. In Section 7.3 we apply these estimates to coordinate charts adapted to the product structure $M = (a, b) \times_f H$ of a Robertson–Walker space–time (cf. Definition 5.10) to study the stability of geodesic incompleteness for such space–times. We show (Theorem 7.19) that if $((a, b) \times_f H, g)$ is a Robertson–Walker space–time with $a > -\infty$, then there is a fine C^0 neighborhood $U(g)$ of g in Lor(M) such that *all* timelike geodesics of *each* space–time (M, g_1) are past incomplete for all $g_1 \in U(g)$. If we assume $b < \infty$ as well, we may obtain (Theorem 7.20) *both* future and past incompleteness of all timelike geodesics for all $g_1 \in U(g)$. A similar result (Theorem 7.23) may be established for null geodesic incompleteness using the C^1 topology on Lor(M). Combining these results yields the C^1 *stability of past nonspacelike geodesic incompleteness for Robertson–Walker space–times.*

In the last section of this chapter we consider space–times which need not have any symmetry properties. We show (Theorem 7.30) that *nonimprisonment is a sufficient condition for the C^1 stability of geodesic incompleteness* [cf. Beem (1994)]. Sufficient conditions for the stability of geodesic completeness involve pseudoconvexity as well as nonimprisonment. Here a space is said to have a *pseudoconvex* class of geodesics if for each compact subset K there is a larger compact subset H such that any geodesic *segment* of the class with endpoints in K lies entirely in H. *Nonimprisonment and pseudoconvexity of nonspacelike geodesics taken together are sufficient for the C^1 stability of nonspacelike geodesic completeness.* This result (Theorem 7.35) implies that nonspacelike geodesic completeness is stable for globally hyperbolic space–times.

At the end of Section 3.6 we discuss the relationship between the choice of warping function $f : (a,b) \to (0,\infty)$ and the nonspacelike geodesic incompleteness of a given Lorentzian warped product space–time $(a,b) \times_f H$.

7.1 Stable Properties of Lor(M) and Con(M)

An equivalence relation C may be placed on the space Lor(M) of Lorentzian metrics for M by defining $g_1, g_2 \in$ Lor(M) to be equivalent if there exists a smooth conformal factor $\Omega : M \to (0,\infty)$ such that $g_1 = \Omega g_2$. As in Chapter 1, we will denote the equivalence class of g in Lor(M) by $C(M,g)$. The quotient space Lor$(M)/C$ of equivalence classes will be denoted by Con(M). There is then a natural projection map τ : Lor$(M) \to$ Con(M) given by $\tau(g) = C(M,g)$.

The fine C^0 topology (cf. Section 3.2) on Lor(M) induces a *quotient topology* on Con(M) as usual. A subset A of Con(M) is defined to be open in this topology if the inverse image $\tau^{-1}(A)$ is open in the fine C^0 topology on Lor(M).

Con(M) may also be given the *interval topology* [cf. Geroch (1970a, p. 447)]. Recall from our discussion of stable causality in Section 3.2 that a partial ordering may be defined on Lor(M) by defining $g_1 < g_2$ if $g_1(v,v) \leq 0$ implies

$g_2(v,v) < 0$ for all $v \neq 0$ in TM. It may then be checked that $g_1, g_2 \in \mathrm{Lor}(M)$ satisfy $g_1 < g_2$ if and only if $g_1' < g_2'$ for all $g_1' \in C(M, g_1)$ and $g_2' \in C(M, g_2)$. Thus the partial ordering $<$ for $\mathrm{Lor}(M)$ projects to a partial ordering on $\mathrm{Con}(M)$ which will also be denoted by $<$. A subbasis for the *interval topology* on $\mathrm{Con}(M)$ is then given by all sets of the form

$$\{C(M,g) \in \mathrm{Con}(M) : C(M,g_1) < C(M,g) < C(M,g_2)\}$$

where g_1 and g_2 are arbitrary Lorentzian metrics for M with $g_1 < g_2$. The quotient and interval topologies agree on $\mathrm{Con}(M)$ [cf. Lerner (1973, p. 23)]. Thus intuitively, two conformal classes $C(M, g_1)$ and $C(M, g_2)$ are close in either of these topologies on $\mathrm{Con}(M)$ if and only if at all points p of M the metrics g_1 and g_2 have light cones which are close in T_pM.

A property defined on $\mathrm{Lor}(M)$ which holds on a C^r-open subset of $\mathrm{Lor}(M)$ is said to be C^r *stable*. Also, a property defined on $\mathrm{Lor}(M)$ which is invariant under the conformal relation C is said to be *conformally stable* if it holds for an open set of equivalence classes in the quotient (or interval) topology on $\mathrm{Con}(M)$. The continuity of the projection map $\tau : \mathrm{Lor}(M) \to \mathrm{Con}(M)$ implies that any conformally stable property defined on $\mathrm{Lor}(M)$ is also C^0 stable on $\mathrm{Lor}(M)$. Furthermore, since the fine C^r topology is strictly finer than the fine C^s topology on $\mathrm{Lor}(M)$ for $r > s$, any conformally stable property defined on $\mathrm{Lor}(M)$ is also C^r stable for all $r \geq 0$.

Example 7.1. (*Stable Causality*) Stable causality is conformally stable and hence also C^r stable for all $r \geq 0$. Indeed, a metric $g \in \mathrm{Lor}(M)$ may be defined to be stably causal if the property of causality is C^0 stable in $\mathrm{Lor}(M)$ at g.

A second example of a conformally stable property is furnished by a result of Geroch (1970a, p. 448).

Theorem 7.2. *Global hyperbolicity is conformally stable and hence C^r stable in* $\mathrm{Lor}(M)$ *for all* $r \geq 0$.

It may also be shown that if S is a smooth Cauchy surface for (M, g), there exists a C^0 neighborhood U of g in $\mathrm{Lor}(M)$ such that if $g_1 \in U$, then S is a Cauchy hypersurface for (M, g_1) [cf. Geroch (1970a, p. 448)].

The Ricci curvature involves the first two partials of the metric tensor but no higher derivatives. Using this fact, Lerner (1973) established the following result.

Proposition 7.3. *If (M, g) is a Lorentzian manifold such that $g(v, v) \leq 0$ and $v \neq 0$ in TM imply $\operatorname{Ric}(g)(v, v) > 0$, then there is a fine C^2 neighborhood $U(g)$ of g in $\operatorname{Lor}(M)$ such that for all $g_1 \in U(g)$, the relations $g_1(v, v) \leq 0$ and $v \neq 0$ in TM imply $\operatorname{Ric}(g_1)(v, v) > 0$.*

It is well known that compact positive definite Riemannian manifolds are always complete. In contrast, compact space–times need not be complete [cf. Fierz and Jost (1965)]. Examples of Williams (1984) show that both geodesic completeness and geodesic incompleteness are, in general, unstable properties for space–times. In fact, the examples show that both of these properties may fail to be stable for compact space–times as well as non-compact space–times. These instabilities are related to questions involving sprays and the Levi–Civita map [cf. Del Riego and Dodson (1988)].

An inextendible closed geodesic c is one which repeatedly retraces the same image. For spacelike and timelike geodesics this implies the geodesic is complete since these geodesics have constant nonzero speed and thus increase in affine parameter by the same amount for each circuit of the image. However, there are closed null geodesics which are inextendible and incomplete. This incompleteness results from the fact that the tangent vector to a null geodesic may fail to return to itself each time the geodesic traverses one circuit of the image. For closed null geodesics, the tangent vector may return to a scalar multiple of itself where the multiple is different from one. In this case, the closed geodesic fails to be a periodic map and the domain is an open subset of $\mathbb{R}$ which is bounded either above or below. More precisely, we have the following result.

Lemma 7.4. *Let (M, g) be a semi-Riemannian manifold with an inextendible null geodesic $\beta : (a, b) \to M$ satisfying $\beta(0) = \beta(1)$ and $\beta'(0) = \lambda\beta'(0)$. If $\lambda = 1$, then β is complete. If $0 < \lambda < 1$, then β is incomplete with $a > -\infty$ and $b = +\infty$. If $1 < \lambda$, then β is incomplete with $a = -\infty$ and $b < +\infty$.*

Proof. If $\lambda = 1$, then $\beta[0,1]$ is a closed geodesic which is periodic (i.e., $\beta(t+1) = \beta(t)$ for all t), and clearly the domain of β is all of $\mathbb{R}$.

If $1 < \lambda$, then the speed of β increases by a factor of λ each time around the image in the positive direction. In particular, if $\gamma(s) = \beta(\lambda s)$, then $\gamma(\lambda^{-1}) = \beta(1) = \beta(0) = \gamma(0)$ and $\gamma'(0) = \lambda\beta'(0) = \beta'(1)$. Thus $\gamma(t) = \beta(1+t)$, and the first circuit of γ (i.e., $0 \leq t \leq \lambda^{-1}$) corresponds to the second circuit of β (i.e., $1 \leq t \leq 1 + \lambda^{-1}$). Thus a countably infinite number of circuits of β in the positive direction starting with $t = 0$ increase the affine parameter of β by

$$\sum_{n=0}^{\infty} \lambda^{-n} = \frac{1}{1-\lambda^{-1}}.$$

It follows that the domain of β has a finite upper bound, and one finds $b = (1-\lambda^{-1})^{-1} < +\infty$. Of course, traversing the geodesic β in the negative direction starting at $t = 0$ changes the affine parameter by $\sum_{n=1}^{\infty} \lambda^n = \infty$, which yields $a = -\infty$.

If $0 < \lambda < 1$, then similar arguments show $a > -\infty$ and $b = +\infty$. $\square$

We will now consider two examples of Williams (1984). These examples are for metrics of the form $dx\,dy + f(x)dy^2$ on the torus. Let $S^1 \times S^1 = M = \{(x,y) \,|\, 0 \leq x \leq 2\pi,\ 0 \leq y \leq 2\pi\}$ with the usual identifications. Let $f : \mathbb{R}^1 \to \mathbb{R}^1$ be a periodic function with period 2π. One may then easily calculate the Christoffel symbols for the metric $g = dx\,dy + f(x)dy^2$.

Example 7.5. (*Instability of Geodesic Completeness*) Let $g = dx\,dy + f(x)dy^2$. If f is identically zero, one has a flat metric $g = dx\,dy$ on the torus M which is geodesically complete. In particular, the closed null geodesic corresponding to $x = 0$ is complete. On the other hand, using $f_n(x) = (1/n)\sin(x)$ one may directly calculate that the closed null geodesic $x = 0$ of $g_n = dx\,dy + f_n(x)dy^2$ is not complete [cf. Williams (1984)]. This incompleteness results from the fact that the tangent vector to this null geodesic of g_n fails to return to itself each time the geodesic traverses the circle corresponding to $x = 0$. Each time around the circle, the tangent vector returns to a scalar multiple of itself where the multiple is different from one. Lemma 7.4 yields that this null geodesic is incomplete for each g_n. Since each C^r neighborhood of the original flat metric g has metrics g_n for large n, it follows that *geodesic completeness is not stable at the Lorentzian metric g.*

This last example may be generalized [cf. Beem and Ehrlich (1987, p. 328)]. If (M, g) is a geodesically complete semi-Riemannian manifold which contains a closed null geodesic, then each C^r neighborhood of g contains incomplete metrics. On the other hand, it has recently been shown that any Lorentzian metric on a compact manifold which has zero curvature, i.e., $R = 0$, is geodesically complete [cf. Carrière (1989), Yurtsever (1992)]. Hence no deformation of the flat metric $g = dx\,dy$ on the torus through *flat* metrics can produce geodesic incompleteness, as in Example 7.5.

Example 7.6. (*Instability of Geodesic Incompleteness*) Williams modified his previous example somewhat to demonstrate the instability of geodesic incompleteness. Using $f(x) = 1 - \cos x$ and x_0 with $0 < x_0 < 2\pi$, one may show that the null geodesic starting at (x_0, y_0) with $dy/dx \neq 0$ at x_0 will be incomplete for the metric $g = dx\,dy + f(x)dy^2$. Williams shows this by solving for $x = x(t)$ and then obtaining an integral formula for $y = y(x)$ involving integrating the reciprocal of $f(x)$. However, he also finds that using $f_n(x) = 1 - \cos x + (1/n)$ one obtains *geodesically complete metrics $g_n = dx\,dy + f_n(x)dy^2$ which are arbitrarily close to the incomplete metric g.*

Examples 7.5 and 7.6 both involve two-dimensional Lorentzian manifolds. However, the examples may be slightly modified to give examples which show geodesic completeness and incompleteness are not necessarily stable for any metric signature (s, r) with both s and r positive.

If (M_1, g_1) and (M_2, g_2) are semi-Riemannian manifolds, the product manifold $M_1 \times M_2$ may be given the usual *product metric* $\overline{g} = g_1 \oplus g_2$, and with this metric the geodesics of the product are of the form $(\gamma_1(t), \gamma_2(t))$ where each factor $\gamma_i(t)$ is either a geodesic of (M_i, g_i) or else is a constant map. Clearly, the product $M_1 \times M_2$ is geodesically complete if and only if each (M_i, g_i) is geodesically complete.

The *semi-Euclidean space* of signature (s, r) is the product manifold $M_1 \times M_2$ where $M_1 = \mathbb{R}^s$ with $g_1 = \sum_{i=1}^{s} -dx_i{}^2$ and $M_2 = \mathbb{R}^r$ with $g_2 = \sum_{i=1}^{r} dx_i{}^2$.

By taking the first manifold (M_1, g_1) to be the manifold (M, g) of Example 7.5 (respectively, Example 7.6) and taking the second manifold (M_2, g_2) to be the semi-Euclidean space of signature $(r-1, s-1)$, one may obtain an example $M_1 \times M_2$ which shows that geodesic completeness (respectively, incompleteness) may fail to be stable for signature (s, r) whenever neither s nor r is zero.

The situation for positive definite (i.e., $s = 0$) and negative definite (i.e., $r = 0$) signatures is quite different. For these signatures, both geodesic completeness and geodesic incompleteness are C^0 stable. For example, if (M, g) is positive definite, then the set $U(g)$ consisting of all positive definite metrics h on M with

$$\frac{1}{4} < \frac{h(v,v)}{g(v,v)} < 4$$

for all nontrivial vectors v is a C^0-open subset of the collection $\mathrm{Riem}(M)$ of all Riemannian metrics on M. Clearly, $U(g)$ contains the metric g. If h is a fixed metric in $U(g)$, then any given curve has an h-length L_h and a g-length L_g with

$$\frac{1}{2} L_g \leq L_h \leq 2L_g.$$

This inequality shows that either g and h are both complete or else both are incomplete. It follows that geodesic completeness and geodesic incompleteness are C^0 stable for positive definite spaces. The same reasoning applies in the negative definite case, and consequently one obtains the following result.

Proposition 7.7. *Let (M, g) be either a positive definite or negative definite semi-Riemannian manifold. If (M, g) is geodesically complete, then for each $r \geq 0$ there is a C^r neighborhood $U(g)$ with each $g_1 \in U(g)$ complete. If (M, g) is geodesically incomplete, then for each $r \geq 0$ there is a C^r neighborhood $U(g)$ with each $g_1 \in U(g)$ incomplete.*

We summarize this last result and the examples of Williams in the following remark.

Remark 7.8. Geodesic completeness and geodesic incompleteness are stable properties in the Whitney C^0 topology for both positive definite and negative definite semi-Riemannian manifolds. For all other metric signatures (s, r), there are examples to show that both geodesic completeness and geodesic incompleteness may fail to be stable for the Whitney C^k topologies for $k \geq 0$.

In this chapter the manifold M is always fixed. However, one-parameter families (M_λ, g_λ) of manifolds and Lorentzian metrics have been considered in general relativity [cf. Geroch (1969)].

7.2 The C^1 Topology and Geodesic Systems

If (M, g) is an arbitrary Lorentzian manifold, then metrics in $\mathrm{Lor}(M)$ which are close to g in the fine C^1 topology have geodesic systems which are close to the geodesic system of g. The purpose of this section is to give a more analytic formulation of this concept, which is needed for our investigation of the C^1 stability of null geodesic incompleteness for Robertson–Walker space–times in Section 7.3.

We begin by recalling a well-known estimate from the theory of ordinary differential equations [cf. Birkhoff and Rota (1969, p. 155)]. We will always use $\|x\|_2$ to denote the Euclidean norm $\left[\sum_{i=1}^{m} {x_i}^2\right]^{1/2}$ of the point $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$.

Proposition 7.9. *Suppose that $f = (f_1, \ldots, f_m)$ and $h = (h_1, \ldots, h_m)$ are continuous functions defined on a common domain $D \subseteq \mathbb{R} \times \mathbb{R}^m$, and suppose that f satisfies the Lipschitz condition*

$$\|f(s,x) - f(s,\overline{x})\|_2 \leq L\|x - \overline{x}\|_2$$

for all $(s,x), (s,\overline{x}) \in D$.

Let $x(s) = (x_1(s), \ldots, x_m(s))$ and $y(s) = (y_1(s), \ldots, y_m(s))$ be solutions for $0 \leq s \leq b$ of the differential equations

$$\frac{dx}{ds} = f(s,x) \qquad \text{and} \qquad \frac{dy}{ds} = h(s,y),$$

respectively. Then if $\|f(s,x) - h(s,x)\|_2 \leq \epsilon$ for all $(s,x) \in D$ with $0 \leq s \leq b$, the following inequality holds for all $0 \leq s \leq b$:

$$\|x(s) - y(s)\|_2 \leq \|x(0) - y(0)\|_2 \cdot e^{Ls} + \frac{\epsilon}{L}(e^{Ls} - 1).$$

Now let M be a smooth manifold, and let (U, x) be any coordinate chart for M. We may obtain an associated coordinate chart

$$\widetilde{x} = (x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_{2n})$$

for $TM|_U$ as follows. Let $\partial/\partial x_1, \ldots, \partial/\partial x_n$ be the basis vector fields defined on U by the local coordinates $x = (x_1, \ldots, x_n)$. Given $v \in T_qM$ for $q \in U$, we may write $v = \sum_{i=1}^n a_i \left.\frac{\partial}{\partial x_i}\right|_q$. Then $\widetilde{x}(v)$ is defined to be $\widetilde{x}(v) = (x_1(q), x_2(q), \ldots, x_n(q), a_1, a_2, \ldots, a_n)$. These coordinate charts may then be used to define Euclidean coordinate distances on U and $TM|_U$. Explicitly, given $p, q \in U$ and $v, w \in TM|_U$, set

$$\|p - q\|_2 = \left(\sum_{i=1}^n [x_i(p) - x_i(q)]^2\right)^{\frac{1}{2}}$$

and

$$\|v - w\|_2 = \left(\sum_{i=1}^{2n} [x_i(v) - x_i(w)]^2\right)^{\frac{1}{2}},$$

respectively. Also if $r \geq 0$ is given, we will use the notation $\|g_1 - g_2\|_{r,U} < \delta$ for $g_1, g_2 \in \mathrm{Lor}(M)$ and a positive constant $\delta > 0$ to mean that calculating with the local coordinates (U, x), all the corresponding entries of the two metric tensors and all their corresponding partial derivatives up to order r are δ-close at each point of U.

We will denote the Christoffel symbols of the second kind for $g_1, g_2 \in \mathrm{Lor}(M)$ by $\Gamma^i_{jk}(g_1)$ and $\Gamma^i_{jk}(g_2)$, respectively. Then for $a = 1, 2$, the geodesic equations in the coordinate chart (U, x) for (M, g_a) are given by

$$\text{(7.1)} \qquad \begin{aligned} \frac{dx_i}{ds} &= x_{i+n}, \\ \frac{dx_{i+n}}{ds} &= -\Gamma^i_{jk}(g_a) x_{j+n} x_{k+n} \end{aligned}$$

for $1 \leq i, j, k \leq n$, where we employ the Einstein summation convention throughout this chapter.

We will use the notation $\exp_q[g_a]$ for the exponential map at $q \in (M, g_a)$, $a = 1, 2$. If $v \in TM|_U$, then $s \to \exp_q[g_a](sv)$ is the solution of (7.1) in U with initial conditions (q, v) for (M, g_a). In order to apply Proposition 7.9 to these exponential maps, we identify $TM|_U$ with a subset of $\mathbb{R}^{2n}$ using the coordinate chart $(TM|_U, \widetilde{x})$ and define $f(s, X) = f(X)$ and $h(s, X) = h(X)$ by

$$\begin{aligned} f_i(X) = h_i(X) &= x_{i+n}, \\ f_{i+n}(X) &= -\Gamma^i_{jk}(g_1)x_{j+n}x_{k+n}, \\ \text{and} \quad h_{i+n}(X) &= -\Gamma^i_{jk}(g_2)x_{j+n}x_{k+n} \end{aligned}$$

for $1 \leq i, j, k \leq n$ and $X = (x_1, \ldots, x_{2n}) \in \mathbb{R}^{2n}$.

Lemma 7.10. *Let (U, x) be a local coordinate chart for the n-manifold M. Let $(p, v) \in TM|_U$, and assume that $c_1(s) = \exp_p[g_1](sv)$ lies in U for all $0 \leq s \leq b$. Given $\epsilon > 0$, there exists a constant $\delta > 0$ such that $\|v - w\|_2 < \delta$ and $\|g_1 - g_2\|_{1,U} < \delta$ imply that $c_2(s) = \exp_q[g_2](sw)$ lies in U for all $0 \leq s \leq b$, and moreover,*

$$|(x_j \circ c_1)(s) - (x_j \circ c_2)(s)| < \epsilon$$

and

$$|(x_j \circ c_1)'(s) - (x_j \circ c_2)'(s)| < \epsilon$$

for all $1 \leq j \leq n$ and $0 \leq s \leq b$.

Proof. Let $f(X)$ and $h(X)$ be defined as above. Then $X(s) = (c_1(s), c_1{}'(s))$ and $Y(s) = (c_2(s), c_2{}'(s))$ are solutions to the differential equations $dX/ds = f(X)$ and $dY/ds = h(Y)$, respectively. Choose D_0 to be an open set in $TM|_U$ about the image of the curve $X(s)$ such that $\overline{D_0}$ is compact. Then there exists a constant L such that f satisfies a Lipschitz condition $\|f(X) - f(\overline{X})\|_2 \leq L\|X - \overline{X}\|_2$ on D_0.

We may make the term $\|X(0) - Y(0)\|_2$ in the estimate of Proposition 7.9 as small as required by making $\|v - w\|_2$ small. Furthermore, since the Christoffel symbols depend only on the coefficients of the metric tensor and on their first partial derivatives, we may make $\|f(X) - h(X)\|_2$ as small as we wish on D_0 by requiring that $\|g_1 - g_2\|_{1,U}$ be small. Hence for a sufficiently small $\delta > 0$, Proposition 7.9 may be applied to guarantee that $c_2(s) \in U$ for all $0 \leq s \leq b$ and also to yield the estimate $\|X(s) - Y(s)\|_2 < \epsilon$ for all $0 \leq s \leq b$. Consequently, $|x_i(s) - y_i(s)| < \epsilon$ for all $1 \leq i \leq 2n$ and $0 \leq s \leq b$. In view of (7.1), this establishes the desired inequalities. □

The following slightly more technical lemma, which is needed in Section 7.3, follows directly from Lemma 7.10 by using the triangle inequality and the continuity of the geodesic solution $X(s) = (c_1(s), {c_1}'(s))$.

Lemma 7.11. *Let (U, x) be a local coordinate chart on the n-manifold M. Suppose that $c_1(v) = \exp_p[g_1](sv)$ lies in U for all $0 \leq s \leq b$. Let $\epsilon > 0$ and s_1 with $0 < s_1 < b$ be given. Then there exists a constant $\delta > 0$ such that if $\|v - w\|_2 < \delta$, $\|g_1 - g_2\|_{1,U} < \delta$, and $|s_0 - s_1| < \delta$, the geodesic $c_2(s) = \exp_q[g_2](sw)$ lies in U for all $0 \leq s \leq b$, and moreover,*

$$|(x_j \circ c_1)(s_1) - (x_j \circ c_2)(s_0)| < \epsilon \tag{7.2}$$

and

$$|(x_j \circ c_1)'(s_1) - (x_j \circ c_2)'(s_0)| < \epsilon \tag{7.3}$$

for all $1 \leq j \leq n$.

Furthermore, if $(x_1 \circ c_1)'(s_1) \neq 0$, then the constant $\delta > 0$ may be chosen such that

$$1 - \epsilon < \left| \frac{(x_1 \circ c_2)'(s_0)}{(x_1 \circ c_1)'(s_1)} \right| < 1 + \epsilon.$$

7.3 Stability of Geodesic Incompleteness for Robertson–Walker Space–times

In this section we investigate the stability in the space of Lorentzian metrics of the nonspacelike geodesic incompleteness of Robertson–Walker space–times $M = (a, b) \times_f H$ (cf. Definition 5.10). It turns out, however, that the proof of the C^0 stability of timelike geodesic incompleteness uses only the homogeneity of the Riemannian factor (H, h) and not the isotropy of (H, h). Accordingly, we will formulate the results in the first portion of this section for the larger class of Lorentzian warped products $M = (a,b) \times_f H$ with $-\infty \leq a < b \leq +\infty$ and (H, h) a homogeneous Riemannian manifold. We will let $x_1 = t$ denote the usual coordinate on (a, b) throughout this section.

In order to study the geodesic incompleteness of such space–times under metric perturbations, it is helpful to use coordinates adapted to the product structure. Fix $p = (t_1, h_1) \in (a,b) \times H$. Since the submanifold $\{t_1\} \times H$ of M is spacelike, the Lorentzian metric g for M restricts to a positive definite inner product on the tangent space to this submanifold at p. Identifying $\{t_1\} \times H$ and H, we may use an orthonormal basis for the tangent space to $\{t_1\} \times H$ at p to define Riemannian normal coordinates $x_2, \ldots, x_n$ for H in a neighborhood V of h_1. Then $(x_1, x_2, \ldots, x_n)$ defines a coordinate system for M on $(a, b) \times V$. By construction, g has the form $\text{diag}\{-1, +1, \ldots, +1\}$ at p in these coordinates. Because the submanifold $\{t_1\} \times H$ is not necessarily totally geodesic if f is nonconstant, these coordinates are not necessarily normal coordinates. Nonetheless, the coordinates $(x_1, x_2, \ldots, x_n)$ are well adapted to the product structure since the level sets $x_1(t) = \lambda$ are just $\{\lambda\} \times V$.

We will say that coordinates $(x_1, \ldots, x_n)$ constructed as above are *adapted at* $p \in M$ and call such coordinates *adapted coordinates*. It will also be useful to define adapted normal neighborhoods.

Definition 7.12. (*Adapted Normal Neighborhood*) An arbitrary convex normal neighborhood U of (M, g) with compact closure $\overline{U}$ is said to be an *adapted normal neighborhood* if $\overline{U}$ is covered by adapted coordinates $(x_1, x_2, \ldots, x_n)$ which are adapted at some point of U such that the following hold:

(1) At every point of U, the components g_{ij} of the metric tensor g expressed in the given coordinates $(x_1, x_2, \ldots, x_n)$ differ from the matrix $\text{diag}\{-1, +1, \ldots, +1\}$ by at most $1/2$.

(2) The metric g satisfies $g <_U \eta_1$, where η_1 is the Minkowskian metric $ds^2 = -2\, dx_1{}^2 + dx_2{}^2 + \cdots + dx_n{}^2$ for U (see the definition of stably causal in Section 3.2 for the notation $g <_U \eta_1$).

Thus on the neighborhood U of Definition 7.12 the Lorentzian metric g may be expressed as

$$g\,|\,U = -dx_1{}^2 + dx_2{}^2 + \cdots + dx_n{}^2 + k_{ij}dx_i dx_j$$

where the functions $k_{ij} : U \to \mathbb{R}$ satisfy $|k_{ij}| \leq 1/2$ for all $1 \leq i, j \leq n$.

For use in the sequel, we need to establish the existence of countable chains $\{U_k\}$ of adapted normal neighborhoods covering future directed, past inextendible, timelike geodesics of the form $c(t) = (t, y_0)$. Since in Definition 7.14 and in Lemma 7.15 it is possible that $a = -\infty$, we adopt the following convention throughout this section.

Convention 7.13. *Let ω_0 denote any fixed interior point of the interval (a, b).*

Now we make the following definition.

Definition 7.14. (*Admissible Chain*) Let $M = (a, b) \times_f H$ denote a Lorentzian warped product with metric $\overline{g} = -dt^2 \oplus fh$. Fix any $y_0 \in H$ and let $c : (a, \omega_0] \to (M, \overline{g})$ be the future directed, past inextendible, timelike geodesic given by $c(t) = (t, y_0)$. A countable covering $\{U_k\}_{k=1}^{\infty}$ of c by open sets and a strictly monotone decreasing sequence $\{t_k\}_{k=1}^{\infty}$ with $t_1 = \omega_0$ and $t_k \to a^+$ as $k \to \infty$ is said to be an *admissible chain* for $c : (a, \omega_0] \to (M, \overline{g})$ if the following two conditions hold:

(1) Each U_k is an adapted normal neighborhood containing $c(t_k) = (t_k, y_0)$ which is adapted at some point of c.

(2) For each k, every future directed and past inextendible nonspacelike curve $\sigma(t) = (t, \sigma_2(t))$ with $\sigma(t_k) = (t_k, y_0)$ remains in U_k for all t with $t_{k+1} \leq t \leq t_k$.

Any future directed nonspacelike curve σ in $(M, \overline{g})$ may be given a parametrization of the form $\sigma(t) = (t, \sigma_1(t))$. Thus condition (2) applies to all future directed nonspacelike curves issuing from (t_k, y_0). We now show that admissible chains exist.

Lemma 7.15. *Let $M = (a, b) \times_f H$ with $a \geq -\infty$ and $\overline{g} = -dt^2 \oplus fh$ be a Lorentzian warped product. For any $y_0 \in H$, the timelike geodesic*

$c : (a, \omega_0] \to (M, \overline{g})$ given by $c(t) = (t, y_0)$ has a covering by an admissible chain.

Proof. We will say that $\{U_k\}$, $\{t_k\}$ is an admissible chain for $c|\,(\theta, \omega_0]$, $\theta \geq a$, if $\{U_k\}$, $\{t_k\}$ satisfy the properties of Definition 7.14 except that $t_k \to \theta^+$ as $k \to \infty$ instead of $t_k \to a^+$ as $k \to \infty$. Set

$$\tau = \inf\{\theta \in [a, \omega_0] : \text{ there is an admissible chain } \{U_k\}, \{t_k\} \text{ for } c|\,(\theta', \omega_0] \text{ for all } \theta' \geq \theta\}.$$

We must show that $\tau = a$.

By taking an adapted normal neighborhood centered at $c(\omega_0)$, it is easily seen that $\tau < \omega_0$. Suppose that $\tau > a$. Let U be any adapted normal neighborhood adapted at the point $(\tau, y_0) \in M$. Choose $r > \tau$ such that all future directed nonspacelike curves $\sigma(t) = (t, \sigma_1(t))$ originating at (r, y_0) lie in U for all $\tau - \epsilon \leq t \leq r$, where $\epsilon > 0$. There exists an admissible chain $\{U_n\}, \{t_n\}$ for $c((r + \tau)/2, \omega_0]$ with $t_m < r$ for some m. Define $\widetilde{U}_m = U_{m+1} = U$. Extending the finite chain $\{U_1, U_2, \ldots, U_{m-1}, \widetilde{U}_m, U_{m+1}\}, \{t_1, t_2, \ldots, t_{m-1}, t_m, \tau - \epsilon\}$ to an infinite admissible chain yields the required contradiction. □

We now show that the subset of U_k for which property (2) of Definition 7.14 holds may be extended from the point (t_k, y_0) to a neighborhood $\{t_k\} \times V_k(y_0)$ in $\{t_k\} \times H$. The notation $\|g - g_1\|_{0,U_k} < \delta$ has been introduced in Section 7.2.

Lemma 7.16. *Let $\{U_k\}, \{t_k\}$ be an admissible chain for the timelike geodesic $c(t) = (t, y_0)$, $c : (a, \omega_0] \to (M, g)$. For each k, there is a neighborhood $V_k(y_0)$ of y_0 in H such that any future directed nonspacelike curve $\sigma(t) = (t, \sigma_1(t))$ with $\sigma(t_k) \in \{t_k\} \times V_k(y_0)$ remains in U_k for all t with $t_{k+1} \leq t \leq t_k$. Furthermore, $V_k(y_0)$ and $\delta > 0$ may be chosen such that if $g_1 \in \mathrm{Lor}(M)$ and $\|g - g_1\|_{0,U_k} < \delta$, then the following two conditions are satisfied. If $\gamma(t) = (t, \gamma_1(t))$ is any nonspacelike curve of (M, g_1) with $\gamma(t_k) \in \{t_k\} \times V_k(y_0)$, then*

(1) *γ remains in U_k for $t_{k+1} \leq t \leq t_k$; and*
(2) *The g_1-length of $\gamma|\,[t_{k+1}, t_k]$ is at most $\sqrt{6}n(t_k - t_{k+1})$.*

Proof. First recall that $\pi : M = (a,b) \times H \to \mathbb{R}$ given by $\pi(t,h) = t$ serves as a global time function for M. In particular, the vector field $\nabla\pi$ satisfies $g(\nabla\pi, \nabla\pi) < 0$ at all points of M. Define $\widetilde{g} \in \mathrm{Lor}(M)$ by

$$\widetilde{g}(x,y) = g(x,y) - g(x,\nabla\pi)\cdot g(y,\nabla\pi).$$

It follows that $g < \widetilde{g}$ on M so that $U_2 = \{\hat{g}_2 \in \mathrm{Con}(M) : \hat{g}_2 < \tau(\widetilde{g})\}$ is an open neighborhood of $C(M,g)$ in $\mathrm{Con}(M)$. Let $U_1 = \tau^{-1}(U_2)$. Then U_1 is a C^0-open neighborhood of g in $\mathrm{Lor}(M)$ such that if $g_1 \in U_1$, the projection map $\pi : M \to \mathbb{R}$ is a global time function for (M, g_1). Hence the hypersurfaces $\{t\} \times H, t \in (a,b)$, remain spacelike in (M, g_1). Thus any nonspacelike curve γ of (M, g_1), $g_1 \in U_1$, may be parametrized as $\gamma(t) = (t, \gamma_1(t))$. Thus the lemma will apply to any nonspacelike curve of (M, g_1) originating at any point of $\{t_k\} \times V_k(y_0)$ provided $g_1 \in U_1$ is sufficiently close to g on U_k.

Let $(x_1, \ldots, x_n)$ denote the given adapted coordinate system for the adapted normal neighborhood U_k. In view of condition (2) of Definition 7.12, we may find $\delta_1 > 0$ such that $\|g - g_1\|_{0,U_k} < \delta_1$ implies $g_1 < \eta_2$ on U_k, where η_2 is the Lorentzian metric on U_k given in the adapted local coordinates by $\eta_2 = -3\,dx_1{}^2 + dx_2{}^2 + \cdots + dx_n{}^2$. Secondly, since C^0-close metrics have close light cones, it follows by a compactness argument that there exist a neighborhood $V_k(y_0)$ of y_0 in H and a constant $\delta_2 > 0$ such that if $g_1 \in \mathrm{Lor}(M)$ satisfies $\|g - g_1\|_{0,U_k} < \delta_2$ and $\gamma(t) = (t, \gamma_1(t))$ is any future directed nonspacelike curve of (M, g_1) with $\gamma(t_k) \in \{t_k\} \times V_k(y_0)$, then $\gamma(t) \in U_k$ for $t_{k+1} \le t \le t_k$.

It remains to establish the length estimate (2). Set $\delta = \min\{\delta_1, \delta_2, 1/2\}$. Suppose that $g' \in \mathrm{Lor}(M)$ satisfies $\|g' - g\|_{0,U_k} < \delta$, and let $\gamma(t) = (t, \gamma_1(t))$ be any nonspacelike curve of (M, g') with $\gamma(t_k) \in \{t_k\} \times V_k(y_0)$, $\gamma : [t_{k+1}, t_k] \to M$. Let $L(\gamma)$ denote the length of γ in (M, g'). Thus

$$L(\gamma) = \int_{t_{k+1}}^{t_k} \sqrt{\sum_{i,j} -g'_{ij}(\gamma(t)) \cdot \gamma_i{}'(t)\,\gamma_j{}'(t)}\; dt.$$

From Definition 7.12 and the choice of the δ's, we have $|g'_{ij}| \le (1 + 1/2) + 1/2 = 2$ and $|\gamma_i{}'(t)| \le \sqrt{3}$ for all $1 \le i, j \le n$. Thus, as required,

$$L(\gamma) \le \int_{t_{k+1}}^{t_k} \sqrt{2n^2(\sqrt{3})^2}\; dt = \sqrt{6}n\,(t_k - t_{k+1}). \qquad \square$$

Assuming now that the Riemannian factor (H, h) of the Lorentzian warped product is homogeneous, we may extend Lemma 7.16 from U_k to $[t_{k+1}, t_k] \times H$. We will use the notation $|g_1 - g|_0 < \delta$, as defined in Section 3.2.

Lemma 7.17. *Let (M, g) be a Lorentzian warped product with (H, h) homogeneous, and let $\{U_k\}, \{t_k\}$ be an admissible chain for $c(t) = (t, y_0)$, $c : (a, \omega_0] \to M$. For each k, there is a continuous function $\delta_k : [t_{k+1}, t_k] \times H \to (0, \infty)$ such that if $g_1 \in \text{Lor}(M)$ and $|g - g_1|_0 < \delta_k$ on $[t_{k+1}, t_k] \times H$, then any nonspacelike curve $\gamma(t) = (t, \gamma_1(t))$, $\gamma : [t_{k+1}, t_k] \to (M, g_1)$, joining any point of $\{t_{k+1}\} \times H$ to any point of $\{t_k\} \times H$, has length at most $\sqrt{6}n(t_k - t_{k+1})$.*

Proof. Fix any $k > 0$. Let $\delta > 0$ be the constant given by Lemma 7.16 such that if $g_1 \in \text{Lor}(M)$ satisfies $\|g_1 - g\|_{0,U_k} < \delta$, then any nonspacelike curve $\gamma(t) = (t, \gamma_1(t))$ in (M, g_1) with $\gamma(t_k) \in \{t_k\} \times V_k(y_0)$ remains in U_k for $t_{k+1} \leq t \leq t_k$ and has length at most $\sqrt{6}n\,(t_k - t_{k+1})$. Also let $(x_1, \ldots, x_n)$ denote the given adapted coordinates for U_k.

We may find isometries $\{\phi_i\}_{i=1}^{\infty}$ in $I(H)$ such that if $y_i = \phi_i(y_0)$ and $V_k(y_i) = \phi_i(V_k(y_o))$, then the sets $\{V_k(y_i)\}_{i=1}^{\infty}$ together with $V_k(y_0)$ form a locally finite covering of H. Let $\Phi_i : M \to M$ be the isometry given by $\Phi_i(t, h) = (t, \phi_i(h))$, and set $\widetilde{U}_i = \Phi_i(U_k)$ for each i. Then the sets $\{\widetilde{U}_i\}$ cover $[t_{k+1}, t_k] \times H$ and $(x_1, x_2 \circ \Phi_i^{-1}, \ldots, x_n \circ \Phi_i^{-1})$ form adapted local coordinates for $\widetilde{U}_i$ for each i. Since everything is constructed with isometries, the constant $\delta > 0$ that works in Lemma 7.16 for U_k and $c(t) = (t, y_0)$ works equally well for each $\widetilde{U}_i$ and $\Phi_i \circ c$, provided that the adapted coordinates $(x_1, x_2 \circ \Phi_i^{-1}, \ldots, x_n \circ \Phi_i^{-1})$ are used for $\widetilde{U}_i$. If we let $\delta_k : [t_{k+1}, t_k] \times H \to M$ be any continuous function such that $\|g_1 - g\|_0 < \delta_k$ on $[t_{k+1}, t_k] \times H$ implies that $\|g_1 - g\|_{0,\widetilde{U}_i} < \delta$ for each i, then the lemma is immediate from Lemma 7.16 $\square$

We are now ready to prove the C^0 *stability of timelike geodesic incompleteness for Lorentzian warped products* $M = (a, b) \times_f H$ with $a > -\infty$ and (H, h) homogeneous.

Theorem 7.18. *Let (M, g) be a warped product space–time of the form $M = (a, b) \times_f H$ with $a > -\infty$, $g = -dt^2 \oplus fh$, and (H, h) a homogeneous*

Riemannian manifold. Then there exists a fine C^0 neighborhood $U(g)$ of g in Lor(M) *of globally hyperbolic metrics such that all timelike geodesics of (M, g_1) are past incomplete for each $g_1 \in U(g)$.*

Proof. Fix any $y_0 \in M$ and let $c : (a, \omega_0] \to M$ be the past inextendible future directed geodesic given by $c(t) = (t, y_0)$. Let $\{U_k\}$, $\{t_k\}$ be an admissible chain for c, guaranteed by Lemma 7.15. Also choose $\delta_k : [t_{k+1}, t_k] \times H \to (0, \infty)$ for each t_k according to Lemma 7.17. Let $\delta : M \to (0, \infty)$ be a continuous function such that $\delta(q) \leq \delta_k(q)$ for all $q \in [t_{k+1}, t_k] \times H$ and each $k > 0$. Set $V_1(g) = \{g_1 \in \text{Lor}(M) : |g_1 - g|_0 < \delta\}$. Since global hyperbolicity is a C^0-open condition, we may also assume that all metrics in $V_1(g)$ are globally hyperbolic.

By the first paragraph of the proof of Lemma 7.16, we may choose a C^0 neighborhood $V_2(g)$ of g in Lor(M) such that for all $g_1 \in V_2(g)$, each hypersurface $\{t\} \times H$, $t \in (a, b)$, is spacelike in (M, g_1). Then every nonspacelike curve $\gamma : (\alpha, \beta) \to (M, g_1)$ may be parametrized in the form $\gamma(t) = (t, \gamma_1(t))$. Hence Lemma 7.16 may be applied to all inextendible nonspacelike geodesics in (M, g_1) with $g_1 \in V_2(g)$.

Now $U(g) = V_1(g) \cap V_2(g)$ is a fine C^0 neighborhood of g in the C^0 topology. Let $g_1 \in U(g)$, and let $\gamma : (\alpha, \beta) \to (M, g_1)$ be any future directed inextendible timelike geodesic. We may assume that $\{t_1\} \times H$ is a Cauchy surface for (M, g_1) by the arguments of Geroch (1970a, p. 448), and hence there exists an $s_0 \in (\alpha, \beta)$ such that $\gamma(s_0) \in \{t_1\} \times H$. In passing from $\{t_{k+1}\} \times H$ to $\{t_k\} \times H$, the g_1 length of γ is at most $\sqrt{6}n\,(t_k - t_{k+1})$, applying Lemma 7.17 for each k. Summing up these estimates, it follows that the g_1 length of $\gamma|\,(\alpha, s_0]$ is at most $\sqrt{6}n\,(t_1 - a)$. Since $\gamma|\,(\alpha, s_0]$ is a past inextendible timelike geodesic of finite g_1 length, it follows that γ is past incomplete in (M, g_1). □

Lerner raised the following question (1973, p. 35) about the Robertson–Walker big bang models (M, g): under small C^2 perturbations of the metric, does each nonspacelike geodesic remain incomplete? Since the Riemannian factor (H, h) of a Robertson–Walker space–time is homogeneous, we obtain the following corollary to Theorem 7.18 which settles affirmatively for timelike geodesics the question raised by Lerner (1973, p. 35).

Theorem 7.19. *Let (M, g) be a Robertson–Walker space–time of the form $M = (a, b) \times_f H$ with $a > -\infty$. Then there exists a fine C^0 neighborhood $U(g)$ of g in* $\mathrm{Lor}(M)$ *of globally hyperbolic metrics such that all timelike geodesics of (M, g_1) are past incomplete for each $g_1 \in U(g)$.*

If we change the time function on (M, g) to $\pi_1 : M \to \mathbb{R}$ defined by $\pi_1(t, h) = -t$, and apply Lemmas 7.16 and 7.17 to the resulting space–time, we obtain the exact analogue of these lemmas for the future directed timelike geodesic $c : [\omega_0, b) \to (M, g)$ given by $c(t) = (t, y_0)$ in the given space–time. Hence if (M, g) is a Lorentzian warped product $M = (a, b) \times_f H$ with (H, h) homogeneous and $b < \infty$, the same proof as for Theorem 7.18 yields the C^0 stability of the future timelike geodesic incompleteness. Combining this remark with Theorem 7.18 then yields the following result.

Theorem 7.20. *Let (M, g) be a Lorentzian warped product of the form $M = (a, b) \times_f H$, $g = -dt^2 \oplus fh$, with both a and b finite and (H, h) homogeneous. Then there is a fine C^0 neighborhood $U(g)$ of g in* $\mathrm{Lor}(M)$ *of globally hyperbolic metrics such that all timelike geodesics of (M, g_1) for each $g_1 \in U(g)$ are both past incomplete and future incomplete.*

It is interesting to note that while the finiteness of a and b is essential to the proof of Theorem 7.20, the proof is independent of the particular choice of warping function $f : (a, b) \to (0, \infty)$. While the homogeneity of the Riemannian factor (H, h) is also used in the proof of Theorem 7.20, no other geometric or topological property of (H, h) is needed.

In general relativity and cosmology, closed big bang models for the universe are considered [cf. Hawking and Ellis (1973, Section 5.3)]. These models are Robertson–Walker space–times for which $b - a < \infty$ and H is compact. Hence Theorem 7.20 implies, in particular, the C^0 stability of timelike geodesic incompleteness for these models.

We now turn to the proof of the C^1 stability of null geodesic incompleteness for Robertson–Walker space–times. Taking $M = (0, 1) \times_f \mathbb{R}$ with $f(t) = (2t)^{-2}$ and $\overline{g} = -dt^2 \oplus f dx^2$, it may be checked using the results of Section 3.6 that the curve $\gamma : (-\infty, 0) \to (M, \overline{g})$ given by $\gamma(t) = (e^t, e^{2t})$ is a past complete null geodesic. Thus by choosing the warping function suitably, it is possible to construct Robertson–Walker space–times with $a > -\infty$ which are past null geodesically complete. Thus unlike the proof of stability for timelike geodesic incompleteness, it is necessary to assume that (M, g) contains a past incomplete (respectively, past and future incomplete) null geodesic to obtain the null analogue of Theorem 7.19 (respectively, Theorem 7.20). Not surprisingly, *the proof of the C^1 stability of null geodesic incompleteness is more complicated than for the timelike case since affine parameters must be used instead of Lorentzian arc length to establish null incompleteness.* Also for the proof of Lemma 7.22, we need the isotropy as well as the homogeneity of (H, h). Thus we will assume that $M = (a, b) \times_f H$ is a Robertson–Walker space–time in the rest of this section.

Let $(V, x_1, \ldots, x_n)$ denote an adapted normal neighborhood of (M, g) with adapted coordinates $(x_1, \ldots, x_n)$. For the proof of Lemma 7.22, it is necessary to define a distance between compact subsets of vectors that are null for different Lorentzian metrics for M and are attached at different points of V. Recall from Section 7.2 that local coordinates $(x_1, \ldots, x_n)$ for V give rise to local coordinates $\widetilde{x} = (x_1, \ldots, x_n, x_{n+1}, \ldots, x_{2n})$ for $TV = TM|_V$. Thus given any $q \in V$, $g_1 \in \text{Lor}(M)$, and $\alpha > 0$, we may define

$$S(q, \alpha, g_1) = \{v \in T_qM : g_1(v, v) = 0 \text{ and } x_{n+1}(v) = -\alpha\}.$$

Then $S(q, \alpha, g_1)$ is a compact subset of T_qM for any $\alpha > 0$ and $g_1 \in \text{Lor}(M)$. Given $p, q \in V$, $g_1, g_2 \in \text{Lor}(M)$, and $\alpha_1, \alpha_2 > 0$, define the *Hausdorff distance* between $S(p, \alpha_1, g_1)$ and $S(q, \alpha_2, g_2)$ by

$$\text{dist}(S(p, \alpha_1, g_1), S(q, \alpha_2, g_2)) =$$
$$\sup_w \inf_v \left\{ \left(\sum_{i=1}^{2n} [x_i(v) - x_i(w)]^2 \right)^{\frac{1}{2}} : \; w \in S(q, \alpha_2, g_2), \; v \in S(p, \alpha_1, g_1) \right\}.$$

The continuity of the components of the metric tensor g as functions $g_{ij} : V \times V \to \mathbb{R}$ and the closeness of light cones for Lorentzian metrics close in the C^0 topology imply the continuity of this distance in p, α, and g [cf. Busemann (1955, pp. 11–12)].

Lemma 7.21. *Let V be an adapted normal neighborhood adapted at $p \in (M, g)$. Given $\alpha > 0$ and $\epsilon > 0$, there exists $\delta > 0$ such that $\|p - q\|_2 < \delta$, $g_1 \in \mathrm{Lor}(M)$ with $\|g - g_1\|_{0,V} < \delta$, and $|\alpha_1 - \alpha| < \delta$ together imply that* $\mathrm{dist}(S(q, \alpha_1, g_1), S(p, \alpha, g)) < \epsilon$.

Now let (M, g) be a Robertson–Walker space–time $(a, b) \times_f H$ which is past null incomplete. Thus some past directed, past inextendible null geodesic $c : [0, A) \to (M, g)$ is past incomplete (i.e., $A < \infty$). Since (H, h) is isotropic and spatially homogeneous, and since isometries map geodesics to geodesics, it follows that all null geodesics are past incomplete. We now fix through the proof of Theorem 7.23 this past inextendible, past incomplete, null geodesic $c : [0, A) \to (M, g)$ with the given parametrization.

Let $(\omega_0, y_0) = c(0) \in M = (a, b) \times_f H$. With this choice of ω_0, apply Lemma 7.15 to the future directed timelike geodesic $t \to (t, y_0)$, $t \leq \omega_0$, to get an admissible chain $\{U_k\}, \{t_k\}$ for this timelike geodesic. Using this choice of $\{t_k\}$, we may find s_k with $0 = s_1 < s_2 < \cdots < s_k < \cdots < A$ such that $c(s_k) \in \{t_k\} \times H$ for each k. Set $\Delta s_k = s_{k+1} - s_k$. As above, let $x_1 : M \to \mathbb{R}$ denote the projection map $x_1(t, h) = t$ on the first factor of $M = (a, b) \times_f H$. Notice that if $(V, x_1, x_2, \ldots, x_n)$ is any adapted coordinate chart, then the coordinate function $x_1 : V \to \mathbb{R}$ coincides with this projection map. If γ is any smooth curve of M which intersects each hypersurface $\{t\} \times H$ of M exactly once and $\gamma(s) \in \{t\} \times H$, we will say that $|(x_1 \circ \gamma)'(s)|$ is the x_1 speed of γ at $\{t\} \times H$. In particular, we will denote by $\alpha_k = |(x_1 \circ c)'(s_k)|$ the x_1 speed of the fixed null geodesic $c : [0, A) \to (M, g)$ at $\{t_k\} \times H$ for each k.

Lemma 7.22. *Let $\epsilon > 0$ be given. Then for each $k > 0$, there exists a continuous function $\delta_k : [t_{k+1}, t_k] \times H \to (0, \infty)$ with the following properties. Let $g_1 \in \mathrm{Lor}(M)$ with $|g - g_1|_1 < \delta_k$ on $[t_{k+1}, t_k] \times H$, and let $\gamma : [0, B) \to M$ be any past directed, past inextendible, null geodesic with $\gamma(0) \in \{t_k\} \times H$ and with x_1 speed of α_k at $\{t_k\} \times H$. Then γ reaches $\{t_{k+1}\} \times H$ with an increase in affine parameter of at most $2\Delta s_k$, and moreover, the x_1 speed θ of γ at $\{t_{k+1}\} \times H$ satisfies the estimate*

$$1 - \epsilon < \left| \frac{\theta}{\alpha_{k+1}} \right| < 1 + \epsilon.$$

Proof. Let $c : [0, A) \to (M, g)$ be the given past incomplete null geodesic as above. Fix $k > 0$. By the spatial homogeneity of Robertson–Walker space-times, we may find an isometry $\phi \in I(H)$ such that $\psi = \text{ id } \times \phi \in I(M, g)$ satisfies $\psi(c(s_k)) = (t_k, y_0)$ with y_0 as above. Since k is fixed during the course of this proof, we may set $p = (t_k, y_0)$ without danger of confusion. Put $c_1(s) = \psi \circ c(s + s_k)$. Then c_1 is a past inextendible, past incomplete, null geodesic of (M, g) with $c_1(0) \in \{t_k\} \times H$, $c_1(\Delta s_k) \in \{t_{k+1}\} \times H$, and $c_1(s) = \exp_p[g](sv)$ for $v = \psi_*(c'(s_k))$. Choose $b > 0$ with $\Delta s_k < b < 2\Delta s_k$ such that $c_1(s) \in U_k$ for all s with $0 \leq s \leq b$. Since $b > \Delta s_k$, we have $c_1(b) \in \{t\} \times H$ for some $t < t_{k+1}$. Hence $(x_1 \circ c_1)(\Delta s_k) - (x_1 \circ c_1)(b) = t_{k+1} - t > 0$. Set $\epsilon_1 = \min\{\epsilon, t_{k+1} - (x_1 \circ c_1)(b)\} > 0$.

Now let $g_1 \in \text{Lor}(M)$, and let $q \in U_k \cap (\{t_k\} \times H)$. Suppose that $\gamma : [0, B) \to (M, g_1)$ is any past directed, past inextendible, null g_1 geodesic with $\gamma(0) = q$ and with x_1 speed α_k at q. Then $w = \gamma'(0) \in T_qM$ satisfies $g_1(w, w) = 0$ and $x_{n+1}(w) < 0$. Moreover, $\gamma(s) = \exp_q[g_1](sw)$. Applying Lemmas 7.10 and 7.11 to c_1 and $c_2 = \gamma$ with the constant ϵ_1 as above, we may find a constant $\delta_0 > 0$ with $0 < \delta_0 < \Delta s_k$ such that $\|v - w\|_2 < \delta_0$, $\|g - g_1\|_{1,U_k} < \delta_0$, and $|s_0 - \Delta s_k| < \delta_0$ imply that

$$(1) \qquad |(x_1 \circ c_1)(s) - (x_1 \circ c_2)(s)| < \epsilon_1 \leq t_{k+1} - (x_1 \circ c_1)(b)$$

for all s with $0 \leq s \leq b$ and

$$(2) \qquad 1 - \epsilon_1 < \left| \frac{(x_1 \circ c_2)'(s_0)}{(x_1 \circ c_1)'(\Delta s_k)} \right| < 1 + \epsilon_1.$$

Setting $s = b$ in (1), we obtain $|(x_1 \circ c_1)(b) - (x_1 \circ c_2)(b)| < t_{k+1} - (x_1 \circ c_1)(b)$, from which $(x_1 \circ c_2)(b) < t_{k+1}$. Hence there exists an s' with $0 < s' < b$ such that $(x_1 \circ c_2)(s') = t_{k+1}$. But then $s' < b < 2\Delta s_k$, which shows that the increase in affine parameter of c_2 in passing from $\{t_k\} \times H$ to $\{t_{k+1}\} \times H$ is less than $2\Delta s_k$ provided that δ_0 is chosen as above.

For any geodesic $c_2(s) = \exp[g_1](sw)$ with g_1 and w δ_0-close to g and v as above, let s' denote the value of the affine parameter s of c_2 such that $x_1 \circ c_2(s') = t_{k+1}$. As $\delta_0 \to 0$, the corresponding value of s' must approach Δs_k by Lemma 7.10. Thus by continuity we may choose δ_1 with $0 < \delta_1 < \delta_0$

such that for any geodesic $c_2(s) = \exp[g_1](sw)$ with g_1 and w both δ_1-close to g and v, we have $x_1 \circ c_2(s') = t_{k+1}$ for some $s' \in [\Delta s_k - \delta_0, \Delta s_k + \delta_0]$. Hence as $\delta_1 < \delta_0$, we may apply estimate (2) above with $s_0 = s'$ to obtain

$$1 - \epsilon < \left| \frac{(x_1 \circ c_2)'(s')}{(x_1 \circ c_1)'(\Delta s_k)} \right| = \left| \frac{\theta}{\alpha_{k+1}} \right| < 1 + \epsilon. \tag{7.4}$$

We now need to extend these estimates from a neighborhood of $v \in T_pM$ to a neighborhood of $S(p, \alpha_k, g)$. To this end, note that since $(a, b) \times_f H = M$ is a warped product of H and the one-dimensional factor (a, b) with $f : (a, b) \to \mathbb{R}$, it follows that $I(H)$ acts transitively on $S(p, \alpha_k, g)$. Thus given any $z \in S(p, \alpha_k, g)$, we may apply the previous arguments using the *same* admissible chain $\{U_k\}$, $\{t_k\}$ to find a constant $\delta_1(z) > 0$ such that if $w \in TM$ satisfies $\|w - z\|_2 < \delta_1(z)$, $\pi(w) \in U_k \cap (\{t_k\} \times H)$, $\|g_1 - g\|_{1,U_k} < \delta_1(z)$, and $c_2(s) = \exp[g_1](sw)$ has x_1 speed α_k at $\{t_k\} \times H$, then c_2 has an increase in affine parameter of at most $2\Delta s_k$ in passing from $\{t_k\} \times H$ to $\{t_{k+1}\} \times H$ and satisfies estimate (7.4). Using the compactness of $S(p, \alpha_k, g)$, we may choose null vectors $v_1, v_2, \ldots, v_j \in S(p, \alpha_k, g)$ such that $S(p, \alpha_k, g)$ is covered by the sets $\{w \in S(p, \alpha_k, g) : \|w - v_m\|_2 < \delta_1(v_m)\}$ for $m = 1, 2, \ldots, j$. Set $\delta_2 = \min\{\delta_1(v_m) : 1 \leq m \leq j\}$. By Lemma 7.21 we may find a constant δ_3 with $0 < \delta_3 < \delta_2$ such that if $\|p - q\|_2 < \delta_3$, $\|g_1 - g\|_{1,U_k} < \delta_3$, and $w \in S(p, \alpha_k, g_1)$, then $\|w - v_m\|_2 < \delta_1(v_m)$ for some m. Hence δ_3 has the following properties. If $\gamma : [0, B) \to (M, g_1)$ is any past inextendible, past directed, null geodesic of (M, g_1) such that $\|g_1 - g\|_{1,U_k} < \delta_3$, $\gamma(0) \in (\{t_k\} \times H) \cap \{q \in U_k : \|p - q\|_2 < \delta_3\}$ where $p = (t_k, y_0)$, and γ has x_1 speed α_k at $\gamma(0)$, then the conclusions of the theorem apply to γ. Since $I(H)$ acts transitively on H, we may extend this result from $(\{t_k\} \times H) \cap \{q \in U_k : \|p - q\|_2 < \delta_3\}$ to all of $\{t_k\} \times H$ just as in the proof of Lemma 7.17. The function $\delta_k : [t_{k+1}, t_k] \times H \to (0, \infty)$ may be constructed exactly as in Lemma 7.17. □

With Lemma 7.22 in hand, we are now ready to prove the C^1 stability of past null geodesic incompleteness for Robertson–Walker space–times. Since Robertson–Walker space–times are isotropic and spatially homogeneous, past incompleteness of one inextendible null geodesic implies past incompleteness of all null geodesics. Thus the stability theorem may be formulated as follows.

Theorem 7.23. *Let (M, g) be a Robertson–Walker space–time containing an inextendible null geodesic which is past incomplete. Then there is a fine C^1 neighborhood $U(g)$ of g in* Lor(M) *of globally hyperbolic metrics such that* all *null geodesics of (M, g_1) are past incomplete for each $g_1 \in U(g)$.*

Proof. Let $M = (a, b) \times_f H$ and let $c : [0, A) \to (M, g)$ be the given inextendible past incomplete null geodesic. With $\omega_0 = x_1(c(0))$, let $\{U_k\}$, $\{t_k\}$, $\{s_k\}$, and $\{\alpha_k\}$ be chosen as in the paragraph preceding Lemma 7.22. Let $\{\beta_k\}$ be a sequence of real numbers with $0 < \beta_k < 1$ for each k such that $1/2 < \prod_{k=1}^{\infty}(1 - \beta_k) < 1$. Thus for each $m \geq 1$, we have

$$1 < \prod_{k=1}^{m}(1 - \beta_k)^{-1} < 2. \tag{7.5}$$

For each $k \geq 1$, we apply Lemma 7.22 with $\epsilon = \beta_k$ to obtain a continuous function $\delta_k : [t_{k+1}, t_k] \times H \to (0, \infty)$ with the properties of Lemma 7.22. Choose a continuous function $\delta : M \to (0, \infty)$ such that for each q in M we have $\delta(q) < \delta_k(q)$ for all k with q in the domain of δ_k. Let $U_1(g) = \{g_1 \in \text{Lor}(M) : |g_1 - g|_1 < \delta\}$. Also choose a C^1-open neighborhood $U_2(g)$ of g in Lor(M) such that all metrics in $U_2(g)$ are globally hyperbolic and such that each hypersurface $\{t\} \times H$ is spacelike in (M, g_1) for all $t \in (a, b)$ and any $g_1 \in U_2(g)$ (cf. Lemma 7.16). Set $U(g) = U_1(g) \cap U_2(g)$.

Now suppose that $g_1 \in U(g)$ and that $\gamma : [0, B) \to M$ is any past directed and past inextendible null geodesic of (M, g_1). Reparametrizing γ if necessary, we may assume that $x_1(\gamma(0)) = t_k$ for some $k \geq 1$ and that γ has x_1 speed α_k at $\{t_k\} \times H$. By Lemma 7.22, γ changes in affine parameter by at most $2\Delta s_k$ in passing from $\{t_k\} \times H$ to $\{t_{k+1}\} \times H$. In order to apply Lemma 7.22 to γ as γ passes from $\{t_{k+1}\} \times H$ to $\{t_{k+2}\} \times H$, it may be necessary to reparametrize γ at $\{t_{k+1}\} \times H$ to have x_1 speed α_{k+1} at $\{t_{k+1}\} \times H$. Nonetheless, if θ_{k+1} denotes the x_1 speed of γ at $\{t_{k+1}\} \times H$, we have $1 - \beta_k < |\theta_{k+1}/\alpha_{k+1}| < 1 + \beta_k$ from Lemma 7.22. Thus the x_1 speed of γ at $\{t_{k+1}\} \times H$ cannot be less than $(1 - \beta_k)\alpha_{k+1}$. Hence the affine parameter of γ increases in passing from $\{t_{k+1}\} \times H$ to $\{t_{k+2}\} \times H$ by at most $2(1 - \beta_k)^{-1}\Delta s_{k+1}$. Arguing inductively, it may be seen that γ increases in affine parameter by at most $2\Delta s_{k+l}\prod_{i=0}^{l-1}(1 - \beta_{k+i})^{-1}$ in passing from $\{t_{k+l}\} \times H$ to $\{t_{k+l+1}\} \times H$.

Using inequality (7.5), we thus have that γ increases in affine parameter by at most $4\Delta s_{k+l}$ in passing from $\{t_{k+l}\}\times H$ to $\{t_{k+l+1}\}\times H$. Since $\sum_{k=1}^{\infty}\Delta s_k = A$, it follows that the total affine length B of γ is less than $4A$. Since $4A < \infty$, it follows that γ is past incomplete as required. □

By reversing the time orientation, we may obtain the analogue of Theorem 7.23 for Robertson–Walker space–times having future incomplete null geodesics. Thus Theorem 7.23 implies the following result.

Theorem 7.24. *Let (M, g) be a Robertson–Walker space–time containing an inextendible null geodesic which is both past and future incomplete. Then there is a fine C^1 neighborhood $U(g)$ of g in* Lor(M) *of globally hyperbolic metrics such that all null geodesics of (M, g_1) are past and future incomplete for each $g_1 \in U(g)$.*

We now obtain two stability theorems for nonspacelike geodesic incompleteness by combining Theorems 7.19 and 7.23 and by combining Theorems 7.20 and 7.24. The first of these theorems applies to all big bang models, and the second theorem applies to the closed big bang models.

Theorem 7.25. *Let (M, g) be a Robertson–Walker space–time of the form $(a, b)\times_f H$, where $a > -\infty$. Assume that (M, g) contains a past incomplete and past inextendible null geodesic. Then there is a fine C^1 neighborhood $U(g)$ of g in* Lor(M) *of globally hyperbolic metrics such that* all *nonspacelike geodesics of (M, g_1) are past incomplete for* each $g_1 \in U(g)$.

Theorem 7.26. *Let (M, g) be a Robertson–Walker space–time of the form $(a, b)\times_f H$, where both a and b are finite. Assume that (M, g) contains an inextendible null geodesic which is both past and future incomplete. Then there is a fine C^1 neighborhood $U(g)$ of g in* Lor(M) *of globally hyperbolic metrics such that* all *nonspacelike geodesics of (M, g_1) are* both *past and future incomplete for* each $g_1 \in U(g)$.

7.4 Sufficient Conditions for Stability

In this section we show that nonimprisonment is a sufficient condition for the C^1 stability of incompleteness [cf. Beem (1994)]. Sufficient conditions for the stability of completeness involve both nonimprisonment and pseudoconvexity. We first state two lemmas which follow from standard results (cf. Lemmas 7.10 and 7.11).

Lemma 7.27. *Let (M, g) be a given semi-Riemannian manifold, fix a geodesic $\gamma : (a, b) \to M$ of (M, g), and let $W = W(\gamma \,|\, [t_1, t_2])$ be a neighborhood of $\gamma \,|\, [t_1, t_2]$. There is a neighborhood V of $\gamma'(t_1)$ in TM and a constant $\epsilon > 0$ such that if $\|g - g_1\|_{1,W} < \epsilon$ and if c is a geodesic of g_1 with $c'(t_1) \in V$, then the domain of c includes the value t_2 and $c(t) \in W$ for all $t_1 \leq t \leq t_2$.*

Lemma 7.28. *Let (M, g) be a given semi-Riemannian manifold, fix a geodesic $\gamma : (a, b) \to M$ of (M, g), and let $W = W(\gamma \,|\, [t_1, t_2])$ be a neighborhood of $\gamma \,|\, [t_1, t_2]$. Given any neighborhood V_1 of $\gamma'(t_1)$ in TM, there is a constant $\epsilon > 0$ and a neighborhood V_2 of $\gamma'(t_2)$ in TM such that if $\|g - g_1\|_{1,W} < \epsilon$ and if c is a geodesic of g_1 with $c'(t_2) \in V_2$, then $c'(t_1) \in V_1$ and $c(t) \in W$ for all $t_1 \leq t \leq t_2$. Furthermore, if γ is timelike (respectively, spacelike), then V_2 and $\epsilon > 0$ may be chosen such that each $v \in \overline{V}_2$ of each such metric g_1 is timelike (respectively, spacelike). If γ is null, then $\epsilon > 0$ may be chosen such that each such metric g_1 has some null vectors in V_2.*

We now define partial imprisonment and imprisonment.

Definition 7.29. (*Imprisonment and Partial Imprisonment*) Let $\gamma : (a, b) \to M$ be an endless (i.e., inextendible) geodesic.

(1) The geodesic γ is *partially imprisoned* as $t \to b$ if there is a compact set $K \subseteq M$ and a sequence $\{t_j\}$ with $t_j \to b^-$ such that $\gamma(t_j) \in K$ for all j.

(2) The geodesic γ is *imprisoned* if there is a compact set K such that the entire image of γ is contained in K.

In other words, γ is partially imprisoned in K as $t \to b$ if either $\gamma(t) \in K$ for all t sufficiently near b or if γ leaves and returns to K an infinite number of times as $t \to b^-$. An imprisoned geodesic is clearly partially imprisoned, but for general space–times one may have some partially imprisoned geodesics which fail to be imprisoned. The stability of incompleteness result (Theorem 7.30) will hold for geodesics which fail to be partially imprisoned in the direction of incompleteness. In contrast, the stability of (nonspacelike) completeness (Theorem 7.35) result will require the nonimprisonment of all nonspacelike geodesics as well as the pseudoconvexity of nonspacelike geodesics.

Let M be given an auxiliary, positive definite, and complete Riemannian metric. This metric has a complete distance function $d_0 : M \times M \to \mathbb{R}$, and the Hopf–Rinow Theorem [cf. Hicks (1965)] guarantees that the compact subsets of M are exactly the subsets which are closed and bounded with respect to d_0. The proof of Theorem 7.30 [cf. Beem (1994)] will use a sequence of n-dimensional annuli $\{W_j\}$ [i.e., sets bounded by spherical shells with respect to the distance d_0 which are centered at a fixed point $\gamma(t_0)$]. The approach involves requiring the perturbed metric g_1 to be sufficiently close to the original metric g on the sequence $\{W_j\}$. For a fixed g_1, one constructs a sequence of geodesics c_j of g_1 such that a given c_j is either tangent or almost tangent to the original γ at a certain point $\gamma(t_j)$ of W_j. A limit geodesic of the sequence will be the desired geodesic c of g_1.

Theorem 7.30. *Let (M, g) be a semi-Riemannian manifold. Assume that (M, g) has an endless geodesic $\gamma : (a, b) \to M$ such that γ is incomplete in the forward direction (i.e., $b \neq \infty$). If γ is not partially imprisoned in any compact set as $t \to b$, then there is a C^1 neighborhood $U(g)$ of g such that each g_1 in $U(g)$ has at least one incomplete geodesic c. Furthermore, if γ is timelike (respectively, null, spacelike) then c may also be taken as timelike (respectively, null, spacelike).*

Proof. Choose some t_0 in the interval (a, b). We will construct sequences t_j, D_j, and L_j. Let t_1 and D_1 be chosen with $D_1 > 1$ and $t_1 > t_0$ such that $d_0(\gamma(t_0), \gamma(t_1)) = D_1$ and $d_0(\gamma(t_0), \gamma(t)) > D_1$ for all $t_1 < t < b$. In other words, $\gamma \,|\, [t_0, b)$ leaves the closed ball of radius D_1 for the last time at t_1. The existence of t_1 follows using the fact that γ is not partially imprisoned as $t \to b$. Set $L_0 = 0$ and $L_1 = 1 + \sup\{d_0(\gamma(t_0), \gamma(t)) \,|\, t_0 < t < t_1\}$. Notice that on the interval $[t_0, t_1]$ the geodesic γ remains within the closed ball of radius $L_1 - 1$ about $\gamma(t_0)$. If $t_1, \ldots, t_{j-1}$, $D_1, \ldots, D_{j-1}$, and $L_0, L_1, \ldots, L_{j-1}$ have been defined, then t_j and D_j are defined by letting $D_j > L_{j-1} + 1$ and requiring both $d_0(\gamma(t_0), \gamma(t_j)) = D_j$ and $d_0(\gamma(t_0), \gamma(t)) > D_j$ for all $t_j < t < b$.

Existence again follows using the nonimprisonment hypotheses. Define $L_j = 1 + \sup\{d_0(\gamma(t_0), \gamma(t)) \mid t_0 < t < t_j\}$. Note that $\lim \gamma(t_j)$ does not exist as $j \to \infty$, and hence $t_j \to b$ as $j \to \infty$. Set $W_1 = \{x \in M \mid d_0(\gamma(t_0), x) < L_1\}$, $W_2 = \{x \in M \mid d_0(\gamma(t_0), x) < L_2\}$, and $W_j = \{x \in M \mid L_{j-2} < d_0(\gamma(t_0), x) < L_j\}$ for $j > 2$. The sets W_j are n-dimensional annuli with respect to the distance d_0. We now define a sequence of positive numbers $\{\epsilon_j\}$ and a sequence of open sets $\{V_j\}$ of TM. Let V_0 be an open neighborhood of $\gamma'(t_0)$ in TM with compact closure, and assume that the closure of V_0 does not contain any trivial vectors. If $g(\gamma'(t_0), \gamma'(t_0)) > 0$ [respectively, $g(\gamma'(t_0), \gamma'(t_0)) < 0$], we assume without loss of generality that all $v \in \overline{V}_0$ satisfy $g(v, v) > 0$ [respectively, $g(v, v) < 0$]. Assume that $V_0, V_1, \ldots, V_{j-1}$ have been defined. Use Lemma 7.28 to obtain an open set $V_j = V_j(\gamma'(t_j))$ of $\gamma'(t_j)$ in TM and a positive number ϵ_j such that if $\|g - g_1\|_{1,P} < \epsilon_j$ where $P = W_j$, and if c is a geodesic of g_1 with $c'(t_j) \in V_j$, then $c'(t_{j-1}) \in V_{j-1}$ and $c(t) \in W_j$ for all $t_{j-1} \le t \le t_j$. Lemma 7.28 implies that if the original geodesic γ is timelike (respectively, spacelike), then we may assume each v in $\overline{V}_j$ is timelike (respectively, spacelike) for g_1. If γ is null, then we may assume that g_1 has some null vectors in V_j. Lemma 7.27 implies we may assume that if c is a geodesic of such a g_1 with $c'(t_j) \in \overline{V}_j$, then the domain of c contains t_{j+1}. We may also assume without loss of generality that the diameter with respect to d_0 of $\pi(V_j)$ is less than $1/2$ and $\epsilon_{j+1} < \epsilon_j$ for all j. Notice that points of M are in at most two consecutive sets of the sequence W_j. It follows that there is a continuous positive-valued function $\epsilon : M \to (0, \infty)$ with $\epsilon(x) < \epsilon_j$ for all j with $x \in W_j$. Let g_1 satisfy $\|g - g_1\|_{1,M} < \epsilon(x)$. We will construct a sequence of geodesics of g_1. Assume first that the original geodesic γ is either timelike or spacelike. For each j, let $c_j(t)$ be the geodesic of g_1 which satisfies ${c_j}'(t_j) = \gamma'(t_j)$. Notice that if γ is timelike (respectively, spacelike), then each c_j is timelike (respectively, spacelike). If the original geodesic γ is null, then we choose c_j such that ${c_j}'(t_j) \in V_j$ and ${c_j}'(t_j)$ is null. The above construction yields that ${c_j}'(t_k) \in V_k$ for all $0 \le k \le j$. Since V_0 has compact closure and this closure does not contain any trivial vectors, one obtains a nontrivial vector v in $\overline{V}_0$ and a subsequence $\{m\}$ of $\{j\}$ such that ${c_m}'(t_0) \to v$. The constructions of V_0 and the sequence $\{c_j\}$ yield that v is timelike (respectively, null or spacelike) for g_1 if the original geodesic γ is timelike (respectively, null or spacelike) for g. Let c be the endless (i.e., inextendible) geodesic of g_1 with $c'(t_0) = v$. Note that each geodesic c_m satisfies $c_m(t_j) \in \pi(V_j)$ for all $0 \leq j \leq m$. Also, for each t in the domain of c, $c_m(t)$ converges to $c(t)$ and $c_m{}'(t)$ converges to $c'(t)$ as $m \to \infty$. By construction, $c'(t_j) \in \overline{V}_j$ implies $c(t_{j+1})$ exists. Furthermore, $c_k{}'(t_{j+1}) \in V_{j+1}$ for all $k > j+1$ yields $c'(t_{j+1}) \in \overline{V}_{j+1}$. Thus, c is defined for all values of t_j, and $d_0(\gamma(t_0), c(t_j)) \to \infty$ as $j \to \infty$. It is easily shown that the limit c is not partially imprisoned in any compact set as $t \to b$. It only remains to show that c cannot have any domain values above b (i.e., $c(b)$ does not exist). Assume the domain of c contains b. Since c is continuous, the set $c([t_0, b])$ would be compact in contradiction to the condition $d_0(\gamma_0(t_0), c(t_j)) \to \infty$. □

An interesting aspect of the construction used in the above proof is that the final geodesic c of the metric g_1 has the same value b for the least upper bound of its domain as the original geodesic γ of the metric g. In essence, this is due to $c_j{}'(t_j) \in V_j$ implying $c_j{}'(t_0) \in V_0$ and to the fact that $d_0(c_j(t_j), c_j(t_0))$ diverges to infinity. Notice that the methods used in the proof of Theorem 7.30 show that if g_1 is another semi-Riemannian metric on M and if g_1 is close to g in the C^1 sense on a neighborhood of γ, then g_1 has a corresponding geodesic which is incomplete. In other words, the proof of Theorem 7.30 only really requires that g and g_1 be metrics which are C^1 close near γ. Thus, if (M, g) has a finite number N of incomplete geodesics and none of these are partially imprisoned in any compact set, then one may construct a C^1 neighborhood $U(g)$ of g such that each g_1 in $U(g)$ has at least N incomplete geodesics. On the other hand, one may construct space–times which have all geodesics complete except for a single geodesic which is incomplete and which is not partially imprisoned [cf. Beem (1976c)]. Thus, the existence of a single incomplete geodesic does not imply that there is more than one such geodesic.

Corollary 7.31. *If (M, g) is a strongly causal space–time which is causally geodesically incomplete, then there is a C^1 neighborhood $U(g)$ of g such that each g_1 in $U(g)$ is nonspacelike geodesically incomplete.*

Proof. This follows from Theorem 7.30 using the fact that for strongly causal space–times, no causal geodesic is future or past partially imprisoned in any compact set. □

It is known that no null geodesic in a two-dimensional Lorentzian manifold has conjugate points. However, a chronological space–time in which all null geodesics have conjugate points is strongly causal and has dimension at least three. Thus, we obtain the following corollary.

Corollary 7.32. *Let (M, g) be a chronological space–time of dimension at least three which is causally geodesically incomplete. If (M, g) has conjugate points along all null geodesics, then there is a C^1 neighborhood $U(g)$ of g such that each g_1 in $U(g)$ is nonspacelike geodesically incomplete.*

Let $(a, b) \times_f H$ be given the metric $g = -dt^2 + f(t)d\sigma^2$. If either a or b is finite, then $(a, b) \times_f H$ is timelike geodesically incomplete. Note that $(a, b) \times_f H$ is necessarily stably causal since $f(t, y) = t$ is a time function. Using Corollary 7.31 and the fact that stably causal space–times are strongly causal, we obtain

Corollary 7.33. *Assume (H, h) is a positive definite Riemannian manifold, and let the warped product $(a, b) \times_f H$ be given the Lorentzian metric $g = -dt^2 \oplus fh$. If either $a \neq -\infty$ or $b \neq +\infty$, then there is a fine C^1 neighborhood $U(g)$ of g such that each metric g_1 in this neighborhood is timelike geodesically incomplete.*

This last corollary applies to a much more general class of space–times than the Robertson–Walker space–times because we have not made symmetry assumptions on (H, h). Of course, for the special case of Robertson–Walker space–times we have already established the stronger conclusion that there is a C^0 neighborhood $U(g)$ with each g_1 in $U(g)$ having all timelike geodesics incomplete (cf. Theorem 7.18).

If K is a subset of $\mathbb{R}^n$, then the *convex hull* K^H of K is the union of all Euclidean line segments with both endpoints in K. It is well known that the convex hull of a compact set is again a compact set. Pseudoconvexity is a generalization of this property to manifolds. In Definition 7.34 below, we assume the class of geodesic segments under consideration to be nonspacelike. One can, of course, consider this property for other classes such as the class of all null geodesic segments, all spacelike segments, etc. Furthermore, pseudoconvexity and disprisonment have also been applied to sprays [cf. Del Riego and Parker (1995)].

Definition 7.34. (*Nonspacelike Pseudoconvexity*) We say (M, g) has a *pseudoconvex nonspacelike geodesic system* if for each compact subset K of M, there is a second compact set H such that each nonspacelike geodesic segment with both endpoints in K lies in H.

The following [cf. Beem and Ehrlich (1987, p. 324)] gives sufficient conditions for the stability of nonspacelike completeness.

Theorem 7.35. *Let (M, g) be a Lorentzian manifold which has no imprisoned nonspacelike geodesics and which has a pseudoconvex nonspacelike geodesic system. If (M, g) is nonspacelike geodesically complete, then there is a C^1 neighborhood $U(g)$ of g in $\mathrm{Lor}(M)$ such that each $g_1 \in U(g)$ is nonspacelike complete.*

Pseudoconvexity is a type of "internal" completeness condition in somewhat the same sense that global hyperbolicity is such a condition. The next proposition shows that nonspacelike pseudoconvexity is a generalization of global hyperbolicity. An example of a nonspacelike pseudoconvex space–time which fails to be globally hyperbolic is given by the open strip $\{(t, x) : 0 < x < 1\}$ in the Minkowski plane.

Proposition 7.36. *If (M, g) is a globally hyperbolic space–time, then (M, g) has a pseudoconvex nonspacelike geodesic system.*

Proof. Let K be a compact subset of M. For each $p \in K$ choose points q and r with q in the chronological past $I^-(p)$ of p and r in the chronological future $I^+(p)$ of p. Then $U(p) = I^+(q) \cap I^-(r)$ is an open set containing p. Since (M, g) is globally hyperbolic, the set $U(p)$ has compact closure given by $J^+(q) \cap J^-(r)$. Cover the compact set K with a finite number of open set $U(p_i) = I^+(q_i) \cap I^-(r_i)$ for $i = 1, 2, \ldots, k$, and let H be the union of the k^2 compact sets of the form $J^+(q_j) \cap J^-(r_i)$ for $1 \leq i, j \leq k$. It is easily seen that all nonspacelike geodesic segments with both endpoints in K must lie in the compact set H. □

Since globally hyperbolic space–times are strongly causal, they have no imprisoned nonspacelike geodesics. Consequently, Theorem 7.35 and Proposition 7.36 yield the following corollary.

Corollary 7.37. *Let (M, g) be a globally hyperbolic space–time. If (M, g) is nonspacelike geodesically complete, then there is a C^1 neighborhood $U(g)$ of g in* $\mathrm{Lor}(M)$ *such that each $g_1 \in U(g)$ is nonspacelike complete.*

Minkowski space–time is globally hyperbolic and geodesically complete. Furthermore, its (entire) geodesic system is pseudoconvex. In other words, given any compact set K, there is a larger compact set H such that any geodesic segment (timelike, null, or spacelike) with endpoints in K must lie in H. In fact, one may take H to be the usual convex hull of K. The next proposition gives the C^1 stability of global hyperbolicity, nonimprisonment, completeness, and inextendibility for Minkowski space–time. The C^1 stability of global hyperbolicity follows from the stronger C^0 stability result of Geroch (1970a). The stability of nonimprisonment and of geodesic completeness (for all geodesics) follows from Beem and Ehrlich (1987, pp. 324–325). See also Beem and Parker (1985, p. 18). The stability of inextendibility follows from the stability of completeness using Proposition 6.16 which guarantees that geodesically complete space–times are inextendible.

Proposition 7.38. *There is a C^1 neighborhood $U(\eta)$ of n-dimensional Minkowski space–time (M, η) such that for each metric g_1 in this neighborhood:*

(1) *(M, g_1) is globally hyperbolic;*
(2) *No geodesic of (M, g_1) is imprisoned;*
(3) *(M, g_1) is geodesically complete; and*
(4) *(M, g_1) is an inextendible space–time.*

CHAPTER 8

MAXIMAL GEODESICS AND CAUSALLY DISCONNECTED SPACE–TIMES

Many basic properties of complete, noncompact Riemannian manifolds stem from the principle that a limit curve of a sequence of minimal geodesics is itself a minimal geodesic. After the correct formulation of completeness had been given by Hopf and Rinow (1931), Rinow (1932) and Myers (1935) were able to establish the existence of a geodesic ray issuing from every point of a complete noncompact Riemannian manifold using this principle. Here a geodesic $\gamma : [0, \infty) \to (N, g_0)$ is said to be a *ray* if γ realizes the Riemannian distance between every pair of its points. Rinow and Myers constructed the desired geodesic ray as follows. Since (N, g_0) is complete and noncompact, there exists an infinite sequence $\{p_n\}$ of points in N such that for every point $p \in N$, $d_0(p, p_n) \to \infty$ as $n \to \infty$. Let γ_n be a minimal (i.e., distance realizing) unit speed geodesic segment from $p = \gamma_n(0)$ to p_n. This segment exists by the completeness of (N, g_0). If $v \in T_pN$ is any accumulation point of the sequence $\{\gamma_n'(0)\}$ of unit tangent vectors in T_pN, then $\gamma(t) = \exp_p tv$ is the required geodesic ray. Intuitively, γ is a ray since it is a limit curve of some subsequence of the minimal geodesic segments $\{\gamma_n\}$. The existence of geodesic rays through every point has been an essential tool in the structure theory of both positively curved [cf. Cheeger and Gromoll (1971, 1972)] and negatively curved [cf. Eberlein and O'Neill (1973)] complete noncompact Riemannian manifolds.

A second application of this basic principle of constructing geodesics as limits of minimal geodesic segments is a concrete geometric realization for complete Riemannian manifolds of the theory of ends for noncompact Hausdorff topological spaces [cf. Cohn–Vossen (1936)]. An infinite sequence $\{p_n\}$ of points in a manifold is said to *diverge to infinity* if, given any compact subset K, only finitely many members of the sequence are contained in K. If a complete Riemannian manifold (N, g_0) has more than one end, there exists a compact subset K of N and sequences $\{p_n\}$ and $\{q_n\}$ which diverge to infinity such that $0 < d_0(p_n, q_n) \to \infty$ and every curve from p_n to q_n meets K for each n. Let γ_n be a minimal (i.e., distance realizing) geodesic segment from p_n to q_n. Since each γ_n meets K, a limit geodesic $\gamma : \mathbb{R} \to M$ may be constructed. Moreover, γ is minimal as a limit of a sequence of minimal curves. Then "$\gamma(-\infty)$" corresponds to the end of N represented by $\{p_n\}$ and "$\gamma(+\infty)$" to the end represented by $\{q_n\}$. In particular, a complete Riemannian manifold with more than one end contains a line, i.e., a geodesic $\gamma : (-\infty, +\infty) \to N$ that is distance realizing between any two of its points.

Motivated by these Riemannian constructions, we study similar existence theorems for geodesic rays and lines in strongly causal space–times. From the viewpoint of general relativity, it is desirable to have constructions that are valid not only for globally hyperbolic subsets of space–times, but also for strongly causal space–times. However, if we only assume strong causality, it is not true in general that causally related points may be joined by maximal geodesic segments. Thus a slightly weaker principle for construction of maximal geodesics is needed for Lorentzian manifolds than for complete Riemannian manifolds. Namely, in strongly causal space–times, limit curves of sequences of "almost maximal" curves are maximal and hence are also geodesics. In Section 8.1 we give two methods for constructing families of almost maximal curves whose limit curves in strongly causal space–times are maximal geodesics. The strong causality is needed to ensure the upper semicontinuity of arc length in the C^0 topology on curves and also so that Proposition 3.34 may be applied. In Section 8.2 we apply this construction to prove the existence of past and future directed nonspacelike geodesic rays issuing from every point of a strongly causal space–time. In Section 8.3 we study the class of causally disconnected space–times. Here a space–time is said to be *causally disconnected by a compact set K* if there are two infinite sequences $\{p_n\}$ and $\{q_n\}$, both diverging to infinity, such that $p_n \leqslant q_n$, $p_n \neq q_n$, and all nonspacelike curves from p_n to q_n meet K for each n. A space–time (M, g) admitting such a compact K causally disconnecting two divergent sequences is said to be *causally disconnected.* It is evident from the definition that causal disconnection is a global conformal invariant of $C(M, g)$. Applying the principle of Section 8.1, we show that if the strongly causal space–time (M, g) is causally disconnected by the compact set K, then (M, g) contains a nonspacelike geodesic line $\gamma : (a, b) \to M$ which intersects K. That is, $d(\gamma(s), \gamma(t)) = L(\gamma \mid [s, t])$ for all s, t with $a < s \leq t < b$. This result is essential to the proof of the singularity theorem 6.3 in Beem and Ehrlich (1979a), as will be seen in Chapter 12. We conclude this chapter by studying conditions on the global geodesic structure of a given space–time (M, g) which imply that (M, g) is causally disconnected. In particular, we show that all two-dimensional globally hyperbolic space–times are causally disconnected. Also, it follows from one of these conditions and the existence of nonspacelike geodesic lines in strongly causal, causally disconnected space–times that a strongly causal space–time containing no future directed null geodesic rays contains a timelike geodesic line.

8.1 Almost Maximal Curves and Maximal Geodesics

The purpose of this section is to show how geodesics may be constructed as limits of "almost maximal" curves in strongly causal space–times. In both constructions, the upper semicontinuity of Lorentzian arc length in the C^0 topology on curves for strongly causal space–times and the lower semicontinuity of Lorentzian distance play important roles. The strong causality of (M, g) is used in our approach here so that convergence in the limit curve sense and in the C^0 topology on curves are closely related (cf. Proposition 3.34). The first construction may be applied to pairs of chronologically related points p, q with $d(p, q) < \infty$. While this approach is therefore sufficient to show the existence of nonspacelike geodesic rays in globally hyperbolic space–times [cf. Beem and Ehrlich (1979c, Theorem 4.2)], it is not valid for points at infinite distance. Accordingly, for use in Sections 8.2 and 8.3, we give a second construction which may be used in arbitrary strongly causal space–times. In Section 14.1 a slightly different approach to constructing maximal segments from sequences of nonspacelike almost maximal curves is presented [cf. Galloway (1986a)]. Here, the use of uniform convergence in a unit speed reparametrization with respect to an auxiliary complete Riemannian metric is employed to dispense with the global requirement of strong causality which we assume in this chapter.

Let (M, g) be an arbitrary space–time and suppose that p and q are distinct points of M with $p \leqslant q$. If $d(p,q) = 0$, then letting γ be any future directed nonspacelike curve from p to q, we have $L(\gamma) \leq d(p,q) = 0$. Hence $L(\gamma) = d(p,q)$ and γ may be reparametrized to a maximal null geodesic segment from p to q by Theorem 4.13. Thus suppose that $p \ll q$, or equivalently, that $d(p,q) > 0$. If $d(p,q) < \infty$ as well, then by Definition 4.1 there exists a future directed nonspacelike curve γ from p to q with

$$d(p,q) \geq L(\gamma) \geq d(p,q) - \epsilon \tag{8.1}$$

for any $\epsilon > 0$. Of course, inequality (8.1) is only a restriction on $L(\gamma)$ provided $\epsilon < d(p,q)$. In this case, we will call γ an *almost maximal* curve.

We note the following elementary consequence of the reverse triangle inequality.

Remark 8.1. Let $\gamma : [0,1] \to M$ be a future directed nonspacelike curve from p to q, $p \neq q$, with

$$d(p,q) - \epsilon \leq L(\gamma) < \infty.$$

Then for any $s < t$ in $[0,1]$, we have

$$L(\gamma|\,[s,t]) \geq d(\gamma(s),\gamma(t)) - \epsilon.$$

Proof. Assume that $L(\gamma|\,[s,t]) < d(\gamma(s),\gamma(t)) - \epsilon$ for some $s < t$ in $[0,1]$. Then

$$\begin{aligned} L(\gamma) &= L(\gamma|\,[0,s]) + L(\gamma|\,[s,t]) + L(\gamma|\,[t,1]) \\ &\leq d(\gamma(0),\gamma(s)) + L(\gamma|\,[s,t]) + d(\gamma(t),\gamma(1)) \\ &< d(p,\gamma(s)) + d(\gamma(s),\gamma(t)) - \epsilon + d(\gamma(t),q) \\ &\leq d(p,q) - \epsilon, \end{aligned}$$

in contradiction. □

We are now ready to give an example of the principle that for strongly causal space–times, limits of almost maximal curves are maximal geodesics. Strong causality is used here (with all curves given a "Lorentzian parametrization") since convergence in the limit curve sense and in the C^0 topology are closely related for strongly causal space–times but not for arbitrary space–times (cf. Section 14.1).

Proposition 8.2. *Let (M, g) be a strongly causal space–time. Suppose that $p_n \to p$ and $q_n \to q$ where $p_n \leqslant q_n$ for each n and $0 < d(p, q) < \infty$. Let $\gamma_n : [a, b] \to M$ be a future directed nonspacelike curve from p_n to q_n with*

$$d(p_n, q_n) \geq L(\gamma_n) \geq d(p_n, q_n) - \epsilon_n > 0 \tag{8.2}$$

where $\epsilon_n \to 0$ as $n \to \infty$. If $\gamma : [a, b] \to M$ is a limit curve of the sequence $\{\gamma_n\}$ with $\gamma(a) = p$ and $\gamma(b) = q$, then $L(\gamma) = d(p, q)$. Thus γ may be reparametrized to be a smooth maximal geodesic from p to q.

Proof. First, γ is nonspacelike by Lemma 3.29. Second, by Proposition 3.34, a subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$ converges to γ in the C^0 topology on curves. By the upper semicontinuity of arc length in this topology for strongly causal space–times (cf. Remark 3.35), we then have

$$\begin{aligned} L(\gamma) &\geq \limsup L(\gamma_m) \\ &\geq \limsup [d(p_m, q_m) - \epsilon_m] \qquad \text{by (8.2)} \\ &\geq d(p, q) \end{aligned}$$

using the lower semicontinuity of Lorentzian distance (Lemma 4.4). But by definition of distance, $d(p, q) \geq L(\gamma)$. Thus $L(\gamma) = d(p, q)$ and the last statement of the proposition follows from Theorem 4.13. □

We now consider a second method for constructing maximal geodesics in strongly causal space–times (M, g) which may be applied to points at infinite Lorentzian distance. For this purpose, we fix throughout the rest of Chapter 8 an arbitrary point $p_0 \in M$ and a complete (positive definite) Riemannian metric h for the paracompact manifold M. Let $d_0 : M \times M \to \mathbb{R}$ denote the

Riemannian distance function induced on M by h. For all positive integers n, the sets

$$\overline{B}_n = \{m \in M : d_0(p_0, m) \leq n\}$$

are compact by the Hopf–Rinow Theorem. Thus the sets $\{\overline{B}_n : n > 0\}$ form a compact exhaustion of M by connected sets. For each n, let

$$d[\overline{B}_n] : \overline{B}_n \times \overline{B}_n \to \mathbb{R} \cup \{\infty\}$$

denote the Lorentzian distance function induced on $\overline{B}_n$ by the inclusion $\overline{B}_n \subseteq (M, g)$. That is, given $p \in \overline{B}_n$, set $d[\overline{B}_n](p, q) = 0$ if $q \notin J^+(p, \overline{B}_n)$, and for $q \in J^+(p, \overline{B}_n)$, let $d[\overline{B}_n](p, q)$ be the supremum of lengths of future directed nonspacelike curves from p to q which are contained in $\overline{B}_n$. It is then immediate that $d[\overline{B}_n](p, q) \leq d(p, q)$ for all $p, q \in \overline{B}_n$. However, $d[\overline{B}_n]$ is *not* in general the restriction of the given Lorentzian distance function d of (M, g) to the set $\overline{B}_n \times \overline{B}_n$. Nonetheless, for strongly causal space–times these two distances coincide "in the limit."

Lemma 8.3. *Let (M, g) be strongly causal. Then for all $p, q \in M$, we have $d(p, q) = \lim d[\overline{B}_n](p, q)$.*

Proof. Since $d[\overline{B}_n](p, q) \leq d(p, q)$, the desired equality is obvious in the event that $d(p, q) = 0$. Thus suppose that $d(p, q) > 0$. By definition of Lorentzian distance, we may find a sequence $\{\gamma_k\}$ of future directed nonspacelike curves from p to q such that $L(\gamma_k) \to d(p, q)$ as $k \to \infty$. (If $d(p, q) = \infty$, choose $\{\gamma_k\}$ such that $L(\gamma_k) \geq k$ for each k.) Since the image of γ_k in M is compact and the Riemannian distance function $d_0 : M \times M \to \mathbb{R}$ is continuous and finite-valued, there exists an $n(k) > 0$ for each k such that $\gamma_k \subseteq \overline{B}_j$ for all $j \geq n(k)$. Thus $d(p, q) = \lim L(\gamma_k) \leq \lim d[\overline{B}_n](p, q)$. Hence as $d[\overline{B}_n](p, q) \leq d(p, q)$ for each n, the lemma is established. □

It will be convenient to introduce the following notational convention for use throughout the rest of this chapter.

Notational Convention 8.4. *Let γ be a future directed nonspacelike curve in a causal space–time. Suppose $p = \gamma(s)$ and $q = \gamma(t)$ with $s < t$ and $p \neq q$. We will let $\gamma[p, q]$ denote the restriction of γ to the interval $[s, t]$.*

For strongly causal space–times, the Lorentzian distance function d and the $d[\overline{B}_n]$'s are related by the following lower semicontinuity.

Lemma 8.5. *Let (M, g) be strongly causal. If $p_n \to p$ and $q_n \to q$, then $d(p,q) \leq \liminf d[\overline{B}_n](p_n, q_n)$.*

Proof. If $d(p,q) = 0$, there is nothing to prove. Thus we first assume that $0 < d(p,q) < \infty$. Let $\epsilon > 0$ be given. By definition of Lorentzian distance and standard results from elementary causality theory [cf. Penrose (1972, pp. 15–16)], a timelike curve γ from p to q may be found with $d(p,q) - \epsilon < L(\gamma) \leq d(p,q)$. Since γ is timelike and $L(\gamma) > d(p,q) - \epsilon$, we may find $r_1, r_2 \in \gamma$ with $d(p,q) - \epsilon < L(\gamma[r_1, r_2])$ and $p \ll r_1 \ll r_2 \ll q$. Since $I^-(r_1)$ and $I^+(r_2)$ are open and $p_n \to p$, $q_n \to q$, we have $p_n \ll r_1 \ll r_2 \ll q_n$ for all n sufficiently large. Also $\gamma \subseteq \overline{B}_n$, $p_n \in J^-(r_1, \overline{B}_n)$, and $q_n \in J^+(r_2, \overline{B}_n)$ for all n sufficiently large. Consequently, $d(p,q) - \epsilon < L(\gamma[r_1, r_2]) \leq d[\overline{B}_n](p_n, q_n)$ for all large n. Since $\epsilon > 0$ was arbitrary, we thus have $d(p,q) \leq \liminf d[\overline{B}_n](p_n, q_n)$ in the case that $0 < d(p,q) < \infty$. Assume finally that $d(p,q) = \infty$. Choosing timelike curves γ_k from p to q with $L(\gamma_k) \geq k$ for each k, we have for each k that $d[\overline{B}_n](p_n, q_n) \geq k - \epsilon$ for all n sufficiently large as above. Hence $\lim d[\overline{B}_n](p_n, q_n) = \infty$ as required. $\square$

Since we are assuming that (M, g) is strongly causal but not necessarily globally hyperbolic, it is possible that the Lorentzian distance function $d : M \times M \to \mathbb{R} \cup \{\infty\}$ assumes the value $+\infty$. Nonetheless, for any given $\overline{B}_n$ the distance function $d[\overline{B}_n] : \overline{B}_n \times \overline{B}_n \to \mathbb{R} \cup \{\infty\}$ is finite-valued. This is a consequence of the compactness of the $\overline{B}_n$ and the compactness of certain subspaces of nonspacelike curves in the C^0 topology on curves [cf. Penrose (1972, p. 50, Theorem 6.5)]. Moreover, this compactness also implies the existence of curves realizing the $d[\overline{B}_n]$ distance for points $p, q \in \overline{B}_n$ with $q \in J^+(p, \overline{B}_n)$.

Lemma 8.6. *Let (M, g) be a strongly causal space–time and let $n > 0$ be arbitrary. If $q \in J^+(p, \overline{B}_n)$, then $d[\overline{B}_n](p,q) < \infty$ and there exists a future directed nonspacelike curve γ in $\overline{B}_n$ joining p to q which satisfies $L(\gamma) = d[\overline{B}_n](p,q)$.*

Proof. By definition of the distance $d[\overline{B}_n]$, if $d[\overline{B}_n](p,q) = 0$ and $q \in J^+(p, \overline{B}_n)$, then there exists a future directed nonspacelike curve γ in $\overline{B}_n$ from p to q with $L(\gamma) \leq d[\overline{B}_n](p,q) = 0$. Hence $L(\gamma) = d[\overline{B}_n](p,q)$ as required. Thus we may suppose that $d[\overline{B}_n](p,q) > 0$. Again by definition of $d[\overline{B}_n]$, we may find a sequence $\{\gamma_k\}$ of future directed nonspacelike curves from p to q with $L(\gamma_k) \to d[\overline{B}_n](p,q)$. (If $d[\overline{B}_n](p,q) = \infty$, choose γ_k with $L(\gamma_k) \geq k$ for each k.) Since $\overline{B}_n$ is compact and (M,g) is strongly causal, there exists a future directed nonspacelike curve γ in $\overline{B}_n$ joining p to q with the property that a subsequence $\{\gamma_m\}$ of $\{\gamma_k\}$ converges to γ in the C^0 topology on curves by Theorem 6.5 of Penrose (1972, pp. 50–51). But then using the upper semicontinuity of arc length in the C^0 topology on curves, we have $d[\overline{B}_n](p,q) = \lim L(\gamma_m) \leq L(\gamma)$ which implies the finiteness of $d[\overline{B}_n](p,q)$. Since $L(\gamma) \leq d[\overline{B}_n](p,q)$ from the definition, we also have $d[\overline{B}_n](p,q) = L(\gamma)$ as required. □

Now let $p, q \in M$ with $p \leqslant q$, $p \neq q$, be arbitrary. Choose any nonspacelike curve γ_0 from p to q. Since the image of γ_0 is compact in M and the Riemannian distance function is continuous, we may find an $N > 0$ such that γ_0 is contained in $\overline{B}_N$. Hence $q \in J^+(p, \overline{B}_n)$ for all $n \geq N$. Thus using Lemma 8.6 we may find a future directed nonspacelike curve γ_n from p to q with $L(\gamma_n) = d[\overline{B}_n](p,q)$ for each $n \geq N$. For C^0 limit curves of the sequence $\{\gamma_n\}$, we then have the following analogue of Proposition 8.2.

Proposition 8.7. *Let (M,g) be strongly causal and let $p, q \in M$ be distinct points with $p \leqslant q$. For all $n > 0$ sufficiently large, let γ_n be a future directed nonspacelike curve from p to q in $\overline{B}_n$ with $L(\gamma_n) = d[\overline{B}_n](p,q)$. If γ is a nonspacelike curve from p to q such that $\{\gamma_n\}$ converges to γ in the C^0 topology on curves, then $L(\gamma) = d(p,q)$ and hence γ may be reparametrized to a maximal geodesic segment from p to q.*

Proof. Using Lemma 8.5 and the upper semicontinuity of arc length in the C^0 topology on curves in strongly causal space–times, we have

$$\begin{aligned} d(p,q) &\le \liminf d\,[\overline{B}_n](p,q) \\ &= \liminf L(\gamma_n) \\ &\le \limsup L(\gamma_n) \le L(\gamma) \\ &\le d(p,q) \end{aligned}$$

as required. □

Now let p, q be distinct points of an arbitrary strongly causal space–time with $p \leqslant q$ and let a sequence $\{\gamma_n\}$ of nonspacelike curves from p to q be chosen as in Proposition 8.7. While a limit curve γ for the sequence $\{\gamma_n\}$ with $\gamma(0) = p$ may always be extracted by Proposition 3.31, we have no guarantee that γ reaches q unless (M,g) is globally hyperbolic. Indeed, if $d(p,q) = \infty$, then there is no maximal geodesic from p to q, and hence no limit curve γ of the sequence $\{\gamma_n\}$ with $\gamma(0) = p$ passes through q. Thus the hypothesis that γ joins p to q in Proposition 8.7 together with the conclusion of Proposition 8.7 implies that $d(p,q) < \infty$. On the other hand, the condition that $d(p,q) < \infty$ does *not* imply that any limit curve γ of $\{\gamma_n\}$ with $\gamma(0) = p$ reaches q when (M,g) is not globally hyperbolic. Examples may easily be constructed by deleting points from Minkowski space–time.

8.2 Nonspacelike Geodesic Rays in Strongly Causal Space–times

The purpose of this section is to establish the existence of past and future directed nonspacelike geodesic rays issuing from every point of a strongly causal space–time (M,g).

Definition 8.8. (*Future and Past Directed Nonspacelike Geodesic Rays*) A *future directed* (respectively, *past directed*) *nonspacelike geodesic ray* is a future (respectively, past) directed, future (respectively, past) inextendible, nonspacelike geodesic $\gamma : [0,a) \to (M,g)$ for which $d(\gamma(0),\gamma(t)) = L(\gamma\,|\,[0,t\,])$ (respectively, $d(\gamma(t),\gamma(0)) = L(\gamma\,|\,[0,t\,])$) for all t with $0 \le t < a$.

The reverse triangle inequality then implies that a nonspacelike geodesic ray is maximal between any pair of its points.

Using Lemmas 8.5 and 8.6 we first prove a proposition that will be needed not only for the proof of the existence of nonspacelike geodesic rays, but also for the proof of the existence of nonspacelike geodesic lines in strongly causal, causally disconnected space–times in Section 8.3. Let $\overline{B}_n$ and $d[\overline{B}_n] : \overline{B}_n \times \overline{B}_n \to \mathbb{R}$ be constructed as in Section 8.1.

Proposition 8.9. *Let (M, g) be a strongly causal space–time and let K be any compact subset of M. Suppose that p and q are distinct points of M such that $p \leqslant q$ and every future directed nonspacelike curve from p to q meets K. Then at least one of the following holds:*

(1) *There exists a future directed maximal nonspacelike geodesic segment from p to q which intersects K.*
(2) *There exists a future directed maximal nonspacelike geodesic which starts at p, intersects K, and is future inextendible.*
(3) *There exists a future directed maximal nonspacelike geodesic which ends at q, intersects K, and is past inextendible.*
(4) *There exists a maximal nonspacelike geodesic which intersects K and is both past and future inextendible.*

Proof. Let γ_0 be any future directed nonspacelike curve in M from p to q. Since $K \cup \gamma_0$ is compact, there exists an $N > 0$ such that $K \cup \gamma_0$ is contained in $\overline{B}_n$ for all $n \geq N$. Hence $q \in J^+(p, \overline{B}_n)$ for all $n \geq N$. Thus by Lemma 8.6, for each $n \geq N$ there exists a future directed nonspacelike curve γ_n in $\overline{B}_n$ joining p to q with $L(\gamma_n) = d[\overline{B}_n](p, q)$. By hypothesis, each γ_n intersects K in some point r_n. Since K is compact, there exists a point $r \in K$ and a subsequence $\{r_m\}$ of $\{r_n\}$ such that $r_m \to r$ as $m \to \infty$. Extend each curve γ_m to a past and future inextendible nonspacelike curve which we will still denote by γ_m. By Proposition 3.31, there exists an inextendible nonspacelike limit curve γ for the subsequence $\{\gamma_m\}$ such that γ contains r. Relabeling if necessary, we may assume that $\{\gamma_m\}$ distinguishes γ.

Now the limit curve γ may contain both p and q, only p, only q, or neither p nor q. These four cases give rise respectively to the four cases (1) to (4) of the proposition. Since the proofs are similar, we will only give the proof for the second case. Thus we assume that $\gamma : (a, b) \to M$ contains $p = \gamma(t_0)$ but not q. We must show that $\gamma |\, [t_0, b)$ is maximal. To this end, let x be an arbitrary point of $\gamma |\, [t_0, b)$. Since $\{\gamma_m\}$ distinguishes γ, we may find points

$x_m \in \gamma_m$ with $x_m \to x$ as $m \to \infty$. Passing to a subsequence $\{\gamma_k\}$ of $\{\gamma_m\}$ if necessary, we may assume by Proposition 3.34 that $\gamma_k[p, x_k]$ converges to $\gamma[p, x]$ in the C^0 topology on curves (recall Notational Convention 8.4). Since $\gamma[p, x]$ is closed in M and $q \notin \gamma$, there exists an open set V containing $\gamma[p, x]$ with $q \notin V$. Since $\gamma_k[p, x_k] \to \gamma[p, x]$ in the C^0 topology on curves, there exists an $N_1 > 0$ such that $\gamma_k[p, x_k] \subseteq V$ for all $k \geq N_1$. Hence $q \notin \gamma_k[p, x_k]$ for all $k \geq N_1$. Thus $\gamma_k[p, x_k] \subseteq \gamma_k[p, q]$ for all $k \geq N_1$ which implies that $L(\gamma_k[p, x_k]) = d[\overline{B}_k](p, x_k)$ for all $k \geq N_1$. By Lemma 8.5 and the upper semicontinuity of arc length in the C^0 topology on curves for strongly causal space–times, we have

$$\begin{aligned} d(p, x) &\leq \liminf d[\overline{B}_k](p, x_k) \\ &= \liminf L(\gamma_k[p, x_k]) \\ &\leq \limsup L(\gamma_k[p, x_k]) \leq L(\gamma[p, x]). \end{aligned}$$

Since $L(\gamma[p, x]) \leq d(p, x)$ by definition of Lorentzian distance, we thus have $d(p, x) = L(\gamma[p, x])$ as required. □

For globally hyperbolic space–times, case (1) of Proposition 8.9 always applies because $J^+(p) \cap J^-(q)$ is compact and no inextendible nonspacelike curve is past or future imprisoned in a compact set. However, space–times which are strongly causal but not globally hyperbolic and which have chronologically related points $p \ll q$ to which exactly one of cases (2) to (4) applies may be constructed by deleting points from Minkowski space–time.

With Proposition 8.9 in hand, we are now ready to prove the existence of past and future directed nonspacelike geodesic rays issuing from each point of a strongly causal space–time. By the usual duality, it suffices to show the existence of a future directed ray at each point.

Theorem 8.10. *Let (M, g) be a strongly causal space–time and let $p \in M$ be arbitrary. Then there exists a future directed nonspacelike geodesic ray $\gamma : [0, a) \to M$ with $\gamma(0) = p$, i.e., $d(p, \gamma(t)) = L(\gamma | [0, t])$ for all t with $0 \leq t < a$.*

Proof. Let $c : [0, b) \to M$ be a future directed, future inextendible, timelike curve with $c(0) = p$. Since (M, g) is strongly causal, c cannot be future imprisoned in any compact set (cf. Proposition 3.13). Thus there is a sequence $\{t_n\}$ with $t_n \to b$ such that $d_0(p, c(t_n)) \to \infty$ as $n \to \infty$. Set $q_n = c(t_n)$ for each n.

We now apply Proposition 8.9 to each pair p, q_n with $K = \{p\}$. Thus for each n, either (1) there is a maximal future directed nonspacelike geodesic segment from p to q_n, or (2) there is a future directed, future-inextendible, nonspacelike geodesic ray starting at p. If case (2) occurs for some n, we are done. Thus assume that for each n there is a maximal future directed nonspacelike geodesic segment γ_n from p to q_n. Extend each γ_n to a future directed, future inextendible, nonspacelike curve, still denoted by γ_n. By Proposition 3.31, the sequence $\{\gamma_n\}$ has a future directed, future inextendible, nonspacelike limit curve $\gamma : [0, a) \to M$ with $\gamma(0) = p$. Relabeling the q_n if necessary, we may suppose that the sequence $\{\gamma_n\}$ itself distinguishes γ.

It remains to show that if $x \in \gamma$ with $p \neq x$ is arbitrary, then $L(\gamma[p, x]) = d(p, x)$. Since $\{\gamma_n\}$ distinguishes γ, we may choose $x_n \in \gamma_n$ for each n such that $x_n \to x$ as $n \to \infty$. Choose a subsequence $\{\gamma_m\}$ of $\{\gamma_n\}$ by Proposition 3.34 such that $\{\gamma_m[p, x_m]\}$ converges to $\gamma[p, x]$ in the C^0 topology on curves. Hence there exists an $N > 0$ such that $\gamma_m[p, x_m] \subseteq \overline{B}_N$ for all $m \geq N$. Since $\overline{B}_N$ is compact and $d_0(p, q_n) \to \infty$, there exists an $N_1 \geq N$ such that $q_m \notin \overline{B}_N$ for all $m \geq N_1$. Hence $L(\gamma_m[p, x_m]) = d(p, x_m) = d\,[\overline{B}_m](p, x_m)$ for all $m \geq N_1$. Using Lemma 8.5 and the upper semicontinuity of Lorentzian arc length in the C^0 topology on curves, we then obtain

$$\begin{aligned} d(p, x) &\leq \liminf d\,[\overline{B}_m](p, x_m) = \liminf L(\gamma_m[p, x_m]) \\ &\leq \limsup L(\gamma_m[p, x_m]) \leq L(\gamma[p, x]) \end{aligned}$$

whence $d(p, x) = L(\gamma[p, x])$ as in Proposition 8.9. □

8.3 Causally Disconnected Space–times and Nonspacelike Geodesic Lines

In this section we define and study the class of causally disconnected space–times. Our definition of this class of space–times is motivated by the geometric realization, discussed in the introduction to this chapter, of the ends of a noncompact complete Riemannian manifold by geodesic lines [cf. Freudenthal (1931) for the original definition of ends of a noncompact Hausdorff topological space]. Recall that an infinite sequence in a noncompact topological space is said to *diverge to infinity* if given any compact subset C, only finitely many elements of the sequence are contained in C.

Definition 8.11. (*Causally Disconnected Space–time*) A space–time is said to be *causally disconnected by a compact set* K if there exist two infinite sequences $\{p_n\}$ and $\{q_n\}$ diverging to infinity such that for each n, $p_n \leqslant q_n$, $p_n \neq q_n$, and all future directed nonspacelike curves from p_n to q_n meet K. A space–time (M, g) that is causally disconnected by some compact K is said to be *causally disconnected.*

Note first that if $k \neq n$, then p_k is not necessarily causally related to q_n or p_n. Also the compact set K may be quite different from a Cauchy surface (cf. Theorem 3.17) and non-globally hyperbolic, strongly causal space–times may be causally disconnected even though they contain no Cauchy surfaces. An example is provided by a Reissner–Nordström space–time with $e^2 = m^2$ (cf. Figure 8.1).

It is immediate from Definition 8.11 that if (M, g) is causally disconnected and $g_1 \in C(M, g)$ is arbitrary, then (M, g_1) is causally disconnected. Thus causal disconnection is a global conformal invariant.

We have previously used a more restrictive version of the concept of causal disconnection in which we assumed in addition to the conditions of Definition 8.11 that $0 < d(p_n, q_n) < \infty$ for each n [cf. Beem and Ehrlich (1979a, p. 171, 1979c)]. With this additional condition our previous definition was, in general, conformally invariant only for the class of globally hyperbolic space–times.

We now give the following definition.

Definition 8.12. (*Nonspacelike Geodesic Line*) Let (M, g) be an arbitrary space–time. A past and future inextendible, future directed, nonspacelike geodesic $\gamma : (a, b) \to M$ is said to be a *nonspacelike geodesic line* if $L(\gamma \mid [s, t]) = d(\gamma(s), \gamma(t))$ for all s, t with $a < s \leq t < b$.

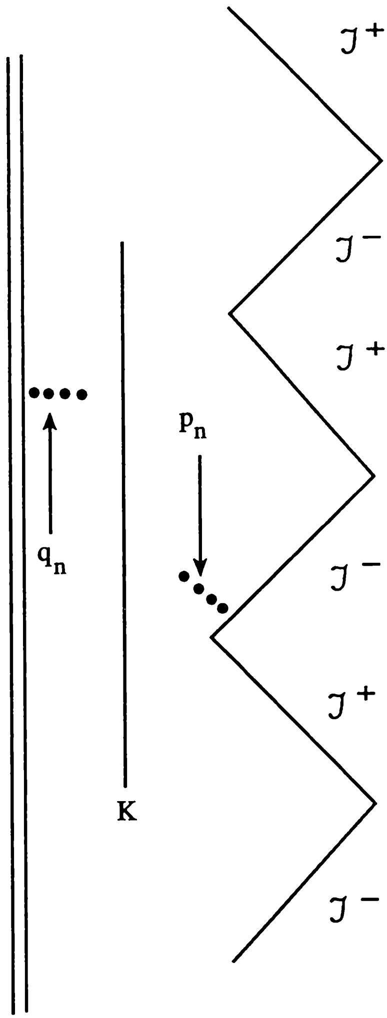


FIGURE 8.1. A Penrose diagram for a Reissner–Nordström space–time with $e^2 = m^2$ containing a causally disconnected set K and associated divergent sequences $\{p_n\}$ and $\{q_n\}$ is shown. This space–time contains no Cauchy surfaces because it is not globally hyperbolic.

We now establish the existence of nonspacelike geodesic lines for strongly causal, causally disconnected space–times. This result will be an important ingredient in the proof of singularity theorems for causally disconnected space–times in Chapter 12.

Theorem 8.13. *Let (M, g) denote a strongly causal space–time which is causally disconnected by a compact set K. Then M contains a nonspacelike geodesic line $\gamma : (a, b) \to M$ which intersects K.*

Proof. Let K, $\{p_n\}$, and $\{q_n\}$ be as in Definition 8.11. Applying Proposition 8.9 to K, p_n, and q_n for each n, we obtain a future directed nonspacelike geodesic γ_n intersecting K at some point r_n and satisfying at least one of the cases (1) to (4) of Proposition 8.9. If case (4) holds for any γ_n, then we are done. Thus assume that no γ_n satisfies case (4). Hence at least one of cases (1), (2), or (3) holds for infinitely many n. Since the proofs are similar, we will only give the proof assuming that case (2) holds for infinitely many n. Passing to a subsequence if necessary, we may suppose that condition (2) holds for all n. Since K is compact, there exists a subsequence $\{r_m\}$ of $\{r_n\}$ such that $r_m \to r$ as $m \to \infty$. Extend each γ_m past p_m to get a nonspacelike curve, still denoted by γ_m, that is past as well as future inextendible for each m. By Proposition 3.31, the sequence $\{\gamma_m\}$ has a future directed, past and future inextendible, nonspacelike limit curve γ with $r \in \gamma$. Relabeling if necessary, we may assume that $\{\gamma_m\}$ distinguishes γ. We will prove that γ is the required nonspacelike line. To prove this, it suffices to show that if $x, y \in \gamma$ are distinct points with $x \leqslant r \leqslant y$, then $L(\gamma[x, y]) = d(x, y)$. Since $\{\gamma_m\}$ distinguishes γ, we may find points $x_m, y_m \in \gamma_m$ such that $x_m \to x$ and $y_m \to y$ as $m \to \infty$. Passing to a subsequence $\{\gamma_k\}$ of $\{\gamma_m\}$ if necessary, we may suppose by Proposition 3.34 that $\gamma_k[x_k, y_k]$ converges to $\gamma[x, y]$ in the C^0 topology on curves. Since $\gamma[x, y]$ is compact in M, there exists an $N > 0$ such that $\gamma[x, y] \subseteq \mathrm{Int}(\overline{B}_N)$. By definition of the C^0 topology on curves, there is then an $N_1 \geq N$ such that $\gamma_k[x_k, y_k] \subseteq \mathrm{Int}(\overline{B}_N)$ for all $k \geq N_1$. Since $\{p_k\}$ diverges to infinity and $\overline{B}_N$ is compact, there is an $N_2 \geq N_1$ such that $p_k \notin \overline{B}_N$ for all $k \geq N_2$. Consequently, x_k comes after p_k on γ_k for all $k \geq N_2$ so that $\gamma_k[x_k, y_k]$ is maximal for all

$k \geq N_2$. We thus have

$$\begin{aligned} d(x,y) &\leq \liminf d(x_k, y_k) = \liminf L(\gamma_k[x_k, y_k]) \\ &\leq \limsup L(\gamma_k[x_k, y_k]) \leq L(\gamma[x,y]) \leq d(x,y). \end{aligned}$$

Hence $d(x,y) = L(\gamma[x,y])$ as required. □

We now give several criteria, expressed in terms of the global geodesic structure, for globally hyperbolic space–times and for strongly causal space–times to be causally disconnected. In particular, we are able to show that all two-dimensional globally hyperbolic space–times are causally disconnected. Also one of our criteria (Proposition 8.18) together with Theorem 8.13 implies that if a strongly causal space–time (M,g) has no null geodesic rays, then (M,g) contains a timelike geodesic line.

Recall that an inextendible null geodesic $\gamma : (a,b) \to (M,g)$ is said to be a *null geodesic line* if $d(\gamma(s), \gamma(t)) = 0$ for all s, t with $a < s \leq t < b$.

Proposition 8.14. *Let (M,g) be strongly causal. If (M,g) contains a null geodesic line, then (M,g) is causally disconnected.*

Proof. Let $c : (a,b) \to (M,g)$ be the given null geodesic line. Choose d with $a < d < b$ and put $K = \{c(d)\}$. Choose sequences $\{s_n\}$, $\{t_n\}$ with $s_n < d < t_n$ and $s_n \to a$, $t_n \to b$. Put $p_n = c(s_n)$ and $q_n = c(t_n)$. Since c is both future and past inextendible and (M,g) is strongly causal, both $\{p_n\}$ and $\{q_n\}$ diverge to infinity by Proposition 3.13. Now because c is a maximal null geodesic, any future directed nonspacelike curve σ from p_n to q_n is a reparametrization of $c|\,[s_n, t_n]$, whence σ meets K as required. (First, σ may be reparametrized to be a smooth future directed null geodesic segment $\sigma : [0,1] \to M$ with $\sigma(0) = p_n$, $\sigma(1) = q_n$ by Theorem 4.13. If $\sigma'(1) \neq \lambda c'(t_n)$ for some $\lambda > 0$, then $q_{n+1} \in I^+(p_n)$. But this contradicts $d(p_n, q_{n+1}) = 0$ since c is maximal. Hence, $\sigma'(1) = \lambda c'(t_n)$ for some $\lambda > 0$. But then since (M,g) is causal, σ must simply be a reparametrization of $c|\,[s_n, t_n]$.) □

Proposition 8.14 implies that Minkowski space–time, de Sitter space–time, and the Friedmann cosmological models are all causally disconnected. Also, the Einstein static universe (cf. Example 5.11) shows that there are globally hyperbolic, causally disconnected space–times which do not have null geodesic lines. Thus the existence of a null geodesic line is *not* a necessary condition for a globally hyperbolic space–time to be causally disconnected.

Evidently, the strong causality in Proposition 8.14 may be replaced by any other nonimprisonment condition which guarantees that both ends of c diverge to infinity, together with the requirement that (M, g) be causal.

In the next proposition we will give a sufficient condition for a strongly causal space–time (M, g) to be causally disconnected. For the proof of this result (Proposition 8.18), it is necessary to recall some additional concepts from elementary causality theory. A subset S of (M, g) is said to be *achronal* if no two points of S are chronologically related. Given a closed subset S of (M, g), the *future Cauchy development* or *domain of dependence* $D^+(S)$ of S is defined as the set of all points q such that every past inextendible nonspacelike curve from q intersects S. The *future Cauchy horizon* $H^+(S)$ is given by $H^+(S) = \overline{D^+(S)} - I^-(D^+(S))$. The *future horismos* $E^+(S)$ of S is defined to be $E^+(S) = J^+(S) - I^+(S)$. An achronal set S is said to be *future trapped* if $E^+(S)$ is compact. Details about these concepts may be found in Hawking and Ellis (1973, pp. 201, 202, 184, and 267, respectively).

For the proof of Proposition 8.18 we also need to use a result first obtained in Hawking and Penrose (1970, p. 537, Lemma 2.12). This result is presented somewhat differently during the course of the proof of Theorem 2 in the text of Hawking and Ellis (1973, p. 266). In the proof of this theorem, it is assumed that dim $M \geq 3$ and that (M, g) has everywhere nonnegative nonspacelike Ricci curvatures and satisfies the generic condition [conditions (1) and (2) of Theorem 2]. However, it may be seen that in the proof of Lemma 8.2.1 and the following corollary in Hawking and Ellis (1973, pp. 267–269), it is only necessary to assume that (M, g) is strongly causal to obtain our Lemma 8.15 and Corollary 8.16. We now state these two results for completeness.

Lemma 8.15. *Let A be a closed subset of the strongly causal space–time (M, g). Then $H^+(\overline{E^+(A)})$ is noncompact or empty.*

From this lemma, one obtains as in Hawking and Penrose (1970, p. 537) or Hawking and Ellis (1973, pp. 268–269) the following corollary.

Corollary 8.16. *Let (M, g) be strongly causal. If S is future trapped in (M, g), i.e., $E^+(S)$ is compact, then there is a future inextendible timelike curve γ contained in $D^+(E^+(S))$.*

It will also be convenient to prove the following lemma for the proof of Proposition 8.18.

Lemma 8.17. *Let (M, g) be strongly causal. If $E^+(p)$ is noncompact, then $E^+(p)$ contains an infinite sequence $\{q_n\}$ which diverges to infinity.*

Proof. If $E^+(p)$ is closed, this is immediate since a closed and noncompact subset of M must be unbounded with respect to d_0. Thus assume that $E^+(p)$ is not closed. Then there exists an infinite sequence $\{x_n\} \subseteq E^+(p)$ such that $x_n \to x \notin E^+(p)$ as $n \to \infty$. Since $x_n \in E^+(p)$, we have $d(p, x_n) = 0$ and hence as $x_n \in J^+(p)$, there exists a maximal future directed null geodesic segment γ_n from p to x_n for each n. Extend each γ_n beyond x_n to a future inextendible nonspacelike curve still denoted by γ_n. By Proposition 3.31, the sequence $\{\gamma_n\}$ has a future inextendible, future directed, nonspacelike limit curve $\gamma : [0, a) \to M$ with $\gamma(0) = p$. We may assume that the sequence $\{\gamma_n\}$ itself distinguishes γ. If $x \in \gamma$, then $x \in J^+(p)$. Since $d(p, x) \leq \liminf d(p, x_n) = 0$, we then have $x \in J^+(p) - I^+(p) = E^+(p)$, in contradiction to the assumption that $x \notin E^+(p)$. Thus $x \notin \gamma$. We now show that $\gamma[0, a)$ is contained in $E^+(p)$. To this end, let $z \in \gamma$ be arbitrary. Since $\{\gamma_n\}$ distinguishes γ, we may find $z_n \in \gamma_n$ such that $z_n \to z$ as $n \to \infty$. By Proposition 3.34, there is a subsequence $\{\gamma_k\}$ of $\{\gamma_n\}$ such that $\gamma_k[p, z_k]$ converges to $\gamma[p, z]$ in the C^0 topology on curves. Since $x \notin \gamma$, we may find an open set U containing $\gamma[p, z]$ such that $x \notin \overline{U}$. Since $x_k \to x$, it follows that z_k comes before x_k on γ_k for all k sufficiently large. Thus $\gamma[p, z_k]$ is maximal and $d(p, z_k) = 0$ for all k sufficiently large. Hence $d(p, z) \leq \liminf d(p, z_k) = 0$. Since z was arbitrary, we have thus shown that $d(p, z) = 0$ for all $z \in \gamma$. Since γ is a nonspacelike curve, γ is then a maximal, future directed, future inextendible geodesic ray. Letting $\{t_n\}$ be any infinite sequence with $t_n \to a^-$ and setting $q_n = \gamma(t_n)$ gives the required divergent sequence. $\square$

With these preliminaries completed, we may now obtain a sufficient condition for strongly causal space–times to be causally disconnected. Minkowski space–time shows that this condition is not a necessary one. Recall that a future directed, future inextendible, null geodesic $\gamma : [0, a) \to M$ is said to be a *null geodesic ray* if $d(\gamma(0), \gamma(t)) = 0$ for all t with $0 \leq t < a$.

Proposition 8.18. *Let (M, g) be strongly causal. If $p \in M$ is not the origin of any future* [*respectively, past*] *directed null geodesic ray, then (M, g) is causally disconnected by the future* [*respectively, past*] *horismos $E^+(p) = J^+(p) - I^+(p)$* [*respectively, $E^-(p) = J^-(p) - I^-(p)$*] *of p.*

Proof. We first show that the assumption that p is not the origin of any future directed null geodesic ray implies that $E^+(p)$ is compact. For suppose that $E^+(p)$ is noncompact. Then there exists an infinite sequence $\{q_n\} \subseteq E^+(p)$ which diverges to infinity by Lemma 8.17. Since $q_n \in E^+(p)$, we have $d(p, q_n) = 0$ for all n. As $q_n \in J^+(p)$, there exists a future directed null geodesic γ_n from p to q_n by Corollary 4.14. Extend each γ_n beyond q_n to a future inextendible nonspacelike curve, still denoted by γ_n. Let γ be a future inextendible nonspacelike limit curve of the sequence $\{\gamma_n\}$ with $\gamma(0) = p$, guaranteed by Proposition 3.31. Using Proposition 3.34 and the fact that the q_n's diverge to infinity, it may be shown along the lines of the proof of Theorem 8.10 that if q is any point on γ with $q \neq p$, $q \geqslant p$, then $L(\gamma\,[p, q]) = d(p, q)$. Thus γ may be reparametrized to a null geodesic ray issuing from p, in contradiction. Hence $E^+(p)$ is compact.

We now show that $E^+(p)$ causally disconnects (M, g). Since $E^+(p)$ is compact, the set $\{p\}$ is future trapped in M. Thus by Corollary 8.16, there is a future inextendible timelike curve γ contained in $D^+(E^+(p))$. Extend γ to a past as well as future inextendible timelike curve, still denoted by γ. From the definition of $D^+(E^+(p))$, the curve γ must meet $E^+(p)$ at some point r. Since $E^+(p)$ is achronal and γ is timelike, γ meets $E^+(p)$ at no other point than r. Now let $\{p_n\}$ and $\{q_n\}$ be two sequences on γ both of which diverge to infinity and which satisfy $p_n \ll r \ll q_n$ for each n (cf. Proposition 3.13). To show that $\{p_n\}$, $\{q_n\}$, and $E^+(p)$ causally disconnect (M, g), we must show that for each n, every nonspacelike curve $\lambda : [0, 1] \to M$ with $\lambda(0) = p_n$ and

$\lambda(1) = q_n$ meets $E^+(p)$. Given λ, extend λ to a past inextendible curve $\widetilde{\lambda}$ by traversing γ up to p_n and then traversing λ from p_n to q_n (cf. Figure 8.2). As $q_n \in D^+(E^+(p))$, the curve $\widetilde{\lambda}$ must intersect $E^+(p)$. Since γ meets $E^+(p)$ only at r, it follows λ intersects $E^+(p)$. Thus $\{p_n\}$, $\{q_n\}$, and $K = E^+(p)$ causally disconnect (M, g) as required. □

Combining Theorem 8.13 and Proposition 8.18, we obtain the following result on the geodesic structure of strongly causal space–times with no null geodesic rays. Examples of such space–times are the Einstein static universes (Example 5.11).

Theorem 8.19. *Let (M, g) be a strongly causal space–time such that no point is the origin of any future directed null geodesic ray. Then (M, g) contains a timelike geodesic line.*

Proof. By Proposition 8.18, (M, g) is causally disconnected. Thus (M, g) contains a nonspacelike line by Theorem 8.13. By hypothesis, the line must be timelike rather than null. □

An equivalent result may be formulated using the hypothesis that no point is the origin of any past directed null geodesic ray.

Using Propositions 8.14 and 8.18, we are now able to show that all two-dimensional globally hyperbolic space–times are causally disconnected. We first establish the following lemma.

Lemma 8.20. *Let (M, g) be a two-dimensional globally hyperbolic space–time. If $E^+(p)$ [respectively, $E^-(p)$] is not compact, then both of the future [respectively, past] directed and future [respectively, past] inextendible null geodesics starting at p are maximal.*

Proof. Assume $E^+(p)$ is noncompact for some $p \in M$. Let $c_1 : [0, a) \to M$ and $c_2 : [0, b) \to M$ be the two future directed, future inextendible, null geodesics with $c_1(0) = c_2(0) = p$. Suppose that c_1 is not maximal. Setting $t_0 = \sup\{t \in [0, a) : d(p, c_1(t)) = 0\}$, we have $0 < t_0 < a$ since c_1 is not maximal and (M, g) is strongly causal (cf. Section 9.2). Let $q = c_1(t_0)$ and choose $t_n \to t_0^+$ with $t_n < a$ for each n. By the lower semicontinuity of Lorentzian

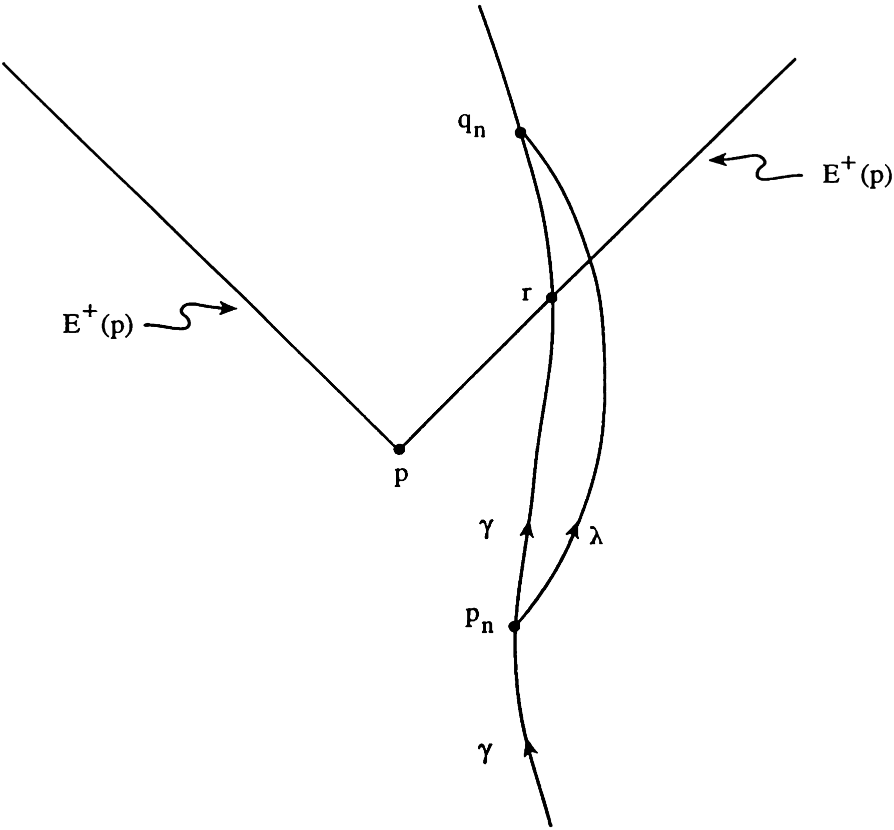


FIGURE 8.2. In the proof of Proposition 8.18, the set $E^+(p)$ is shown to causally disconnect (M, g) if p is not the origin of any future directed null geodesic ray. The timelike curve γ intersects $E^+(p)$ in the single point r, and any nonspacelike curve λ from p_n to q_n must meet $E^+(p)$.

distance, we have $d(p, q) = 0$. Since $c_1(t_n) \in I^+(p)$ and (M, g) is globally hyperbolic, there is a maximal future directed timelike geodesic γ_n from p to

$c_1(t_n)$ for each n. The sequence $\{\gamma_n\}$ has a limit curve γ which is a nonspacelike geodesic, and γ joins p to q by Corollary 3.32. Furthermore, $d(p,q) = 0$ implies that γ is a null geodesic. Since $\dim M = 2$, γ is either a geodesic subsegment of c_1 or a subsegment of c_2. If $\gamma \subseteq c_2$, then c_2 passes through q at some parameter value t_0'. In this case $E^+(p) = \{c_1 \mid [0,t_0]\} \cup \{c_2 \mid [0,t_0']\}$ is compact and we have a contradiction.

Assume $\gamma \subseteq c_1$ and let $U(q)$ be a convex normal neighborhood about q. Let $U(q)$ be chosen so small that no nonspacelike curve which leaves $U(q)$ ever returns. The inextendible null geodesics of $U(q)$ may be divided into two disjoint families $\mathcal{F}_1$ and $\mathcal{F}_2$, each of which cover $U(q)$ simply (cf. Section 3.4). Let $\mathcal{F}_1$ be the class which contains the null geodesic $c_1 \cap U(q)$. Let c_3 be the unique null geodesic of $\mathcal{F}_2$ which contains q. For some fixed large n the timelike geodesic segment γ_n from p to $c_1(t_n)$ must intersect c_3 at some point r. We must have $q \leqslant r$ since if $r \leqslant q$, then $p \ll r$ would yield $p \ll q$, which is false. However, $q \leqslant r$, $q = c_1 \cap c_3$, $q \leqslant c_1(t_n)$, and $r \in c_3$ imply $r \nleqslant c_1(t_n)$, [cf. Busemann and Beem (1966, p. 245)]. This contradicts the fact that $r \ll c_1(t_n)$ because r comes before $c_1(t_n)$ on the timelike geodesic γ_n. This establishes the lemma. □

Theorem 8.21. *All two-dimensional globally hyperbolic space–times are causally disconnected.*

Proof. Let (M,g) be a two-dimensional globally hyperbolic space–time. If $E^+(p_0)$ is compact for some $p_0 \in M$, then each future directed null geodesic starting at p_0 has a cut point and thus fails to be globally maximal (cf. Section 9.2). Thus (M,g) is causally disconnected by Proposition 8.18. Now suppose $E^+(p)$ is noncompact for each $p \in M$. Assume $c : (a,b) \to M$ is a future directed inextendible null geodesic. Let s, t with $a < s \leq t < b$ be arbitrary. Applying Lemma 8.20 at the point $p = c(s)$, we find that $c \mid [s,b)$ is maximal and thus $d(c(s),c(t)) = L(c \mid [s,t])$. This implies that c is maximal and hence c is a null line. Using Proposition 8.14, it follows that (M,g) is causally disconnected. □

We obtain the following corollary to Theorems 8.13 and 8.21.

Corollary 8.22. *Let (M, g) be any globally hyperbolic space–time of dimension $n = 2$. Then (M, g) contains a nonspacelike line.*

CHAPTER 9

THE LORENTZIAN CUT LOCUS

Let $c : [0, \infty) \to N$ be a geodesic in a complete Riemannian manifold starting at $p = c(0)$. Consider the set of all points q on c such that the portion of c from p to q is the unique shortest curve in all of N joining p to q. If this set has a farthest limit point, this limit point is called the *cut point* of p along the ray c. The cut locus $C(p)$ is then defined to be the set of cut points along all geodesic rays starting at p. Since nonhomothetic conformal changes do not preserve pregeodesics, the cut locus of a point in a manifold is *not* a conformal invariant.

The cut locus has played a key role in modern global Riemannian geometry, notably in connection with the Sphere Theorem of Rauch (1951), Klingenberg (1959, 1961), and Berger (1960). The notion of cut point, as opposed to the related but different concept of conjugate point, was first defined by Poincaré (1905). An observation of Poincaré (1905) important in the later work of Rauch, Klingenberg, and Berger was that for a complete Riemannian manifold, if q is on the cut locus of p, then either q is conjugate to p, or else there exist at least two geodesic segments of the same shortest length joining p to q [cf. Whitehead (1935)]. Klingenberg (1959, p. 657) then showed that if q is a closest cut point to p and q is not conjugate to p, there is a geodesic loop at p containing q. Klingenberg used this result to obtain an upper bound for the injectivity radius of a positively curved, complete Riemannian manifold in terms of a lower bound for the sectional curvature and the length of the shortest nontrivial smooth closed geodesic on N.

The importance of the cut locus in Riemannian geometry suggests investigating the analogous concepts and results for timelike and null geodesics in space–times. The central role that conjugate points, which are closely related to cut points, have played in singularity theory in general relativity (cf. Chapter 12) supports this idea. While there are many similarities between the Riemannian cut locus and the locus of timelike cut points in a space–time, there are also striking differences between the Riemannian and Lorentzian cut loci. Most notably, null cut points are invariant under conformal changes. Thus the null cut locus is an invariant of the causal structure of the space–time (M, g). This may be used [cf. Beem and Ehrlich (1979a, Corollary 5.3)] to show that if (M, g) is a Friedmann cosmological model, then there is a C^2 neighborhood $U(g)$ in the space $C(M, g)$ of Lorentzian metrics for M globally conformal to g such that every null geodesic in (M, g_1) is incomplete for all metrics $g_1 \in U(g)$. Because there are intrinsic differences between null and timelike cut points, we prefer to treat these cases separately. One such difference is that unlike null cut points, timelike cut points are not invariant under global conformal changes of Lorentzian metric.

In Section 9.1 we consider the analogue of Riemannian cut points for timelike geodesics. In the Lorentzian setting, timelike geodesics locally maximize the arc length between any two of their points. Thus the appropriate question to ask in defining timelike cut points is whether the portion of the given timelike geodesic segment from p to q is the *longest* nonspacelike curve in *all* of M joining p to q. This may be conveniently formulated using the Lorentzian distance function. Let $\gamma : [0, a) \to M$ be a future directed, future inextendible, timelike geodesic in an arbitrary space–time. Set $t_0 = \sup\{t \in [0, a) : d(\gamma(0), \gamma(t)) = L(\gamma|\, [0, t])\}$. If $0 < t_0 < a$, then $\gamma(t_0)$ is said to be the *future timelike cut point* of $\gamma(0)$ along γ. The future timelike cut point $\gamma(t_0)$ then has the following desired properties: (i) if $t < t_0$, then $\gamma|\, [0, t]$ is the only maximal timelike curve from $\gamma(0)$ to $\gamma(t)$ up to reparametrization; (ii) $\gamma|\, [0, t]$ is maximal for any $t \leq t_0$; (iii) if $t > t_0$, there exists a future directed nonspacelike curve σ from $\gamma(0)$ to $\gamma(t)$ with $L(\sigma) > L(\gamma|\, [0, t])$; and (iv) the future cut point $\gamma(t_0)$ comes at or before the first future conjugate point of $\gamma(0)$ along γ.

Since many of the theorems for Riemannian cut points are true only for *complete* Riemannian manifolds, it is not surprising that the more “global” results given in the rest of Section 9.1 often require global hyperbolicity. What is essential here is to know that chronologically related points at arbitrarily large distances may be joined by maximal timelike geodesic segments. Even so, the proofs are more technical for space–times than for Riemannian manifolds. Instead of using the exponential map directly as in Riemannian geometry, it is necessary to regard a sequence of maximal timelike geodesics as a sequence of nonspacelike curves, extract a limit curve, take a subsequence converging to the limit in the C^0 topology (by Section 3.3), and finally, using the upper semicontinuity of arc length, prove that the limit curve is maximal and hence a geodesic. This technical argument, which is isolated in Lemma 9.6, yields the following analogue of Poincaré's Theorem for complete Riemannian manifolds. If (M, g) is a globally hyperbolic space–time and q is the future cut point of p along the timelike geodesic segment c from p to q, then either one or possibly both of the following hold: (i) q is the first future conjugate point to p; or (ii) there exist at least two maximal geodesic segments from p to q.

In Section 9.2 we study null cut points. Even though null geodesics have zero arc length, null cut points may still be defined using the Lorentzian distance function. Let $\gamma : [0, a) \to M$ be a future directed, future inextendible, null geodesic with $p = \gamma(0)$. Set $t_0 = \sup\{t \in [0, a) : d(p, \gamma(t)) = 0\}$. If $0 < t_0 < a$, then $\gamma(t_0)$ is called the future null cut point of $\gamma(0)$ along γ. The null cut point, if it exists, has the following properties: (i) γ is maximizing up to and including the null cut point; (ii) there is no timelike curve joining p to $\gamma(t)$ for any $t \leq t_0$; (iii) if $t_0 < t < a$, there is a timelike curve from $\gamma(0)$ to $\gamma(t)$; and (iv) the future null cut point comes at or before the first future conjugate point of $\gamma(0)$ along γ. For globally hyperbolic space–times, it is true for null as well as timelike cut points that the analogue of Poincaré's Theorem for complete Riemannian manifolds is valid. Thus if (M, g) is globally hyperbolic and q is the future null cut point of $p = \gamma(0)$ along the null geodesic γ, then either one or possibly both of the following hold: (i) q is the first future conjugate point of p along γ; or (ii) there exist at least two maximal null geodesic segments from p to q. We conclude Section 9.2 by using null cut points to prove singularity theorems for null geodesics following Beem and Ehrlich (1979a, Section 5).

The nonspacelike cut locus, the union of the null and timelike cut loci of a given point, is studied in Section 9.3. For complete Riemannian manifolds, if q is a closest cut point to p, then either q is conjugate to p or there is a geodesic loop based at p passing through q. The globally hyperbolic analogue (Theorem 9.24) of this result has a slightly different flavor, however. If (M, g) is a globally hyperbolic space–time and $q \in M$ is a closest (nonspacelike) cut point to p, then q is either conjugate to p or else q is a *null* cut point to p. Thus there is no closest nonconjugate timelike cut point to p. We also show that for globally hyperbolic space–times, the nonspacelike and null cut loci are closed (Proposition 9.29). It can be seen (Example 9.28) that the hypothesis of global hyperbolicity is necessary here.

In Section 9.4 we treat null and timelike cut points simultaneously, but nonintrinsically, using a different tool than the unit timelike sphere bundle of Section 9.1. This approach, developed in collaboration with G. Galloway, while non-intrinsic, enables the hypothesis of timelike geodesic completeness to be deleted from Proposition 9.30.

9.1 The Timelike Cut Locus

Recall that a future directed nonspacelike curve γ from p to q is said to be maximal if $d(p,q) = L(\gamma)$. We saw above (Theorem 4.13) that a maximal future directed nonspacelike curve may be reparametrized to be a geodesic. We also recall the following analogue of a classical result from Riemannian geometry. The proof may be given along the lines of Kobayashi (1967, p. 99), using in place of the minimal geodesic segment from p_1 to p_2 in the Riemannian proof the fact that if $p \ll q$ and p and q are contained in a convex normal neighborhood, then p and q may be joined by a maximal timelike geodesic segment which lies in this neighborhood.

Lemma 9.1. *Let $c : [0, a] \to M$ be a maximal timelike geodesic segment. Then for any s, t with $0 \leq s < t < a$, the curve $c \mid [s, t]$ is the unique maximal geodesic segment (up to parametrization) from $c(s)$ to $c(t)$.*

Before commencing our study of the timelike cut locus, we need to define the unit future observer bundle $T_{-1}M$ [cf. Thorpe (1977a,b)].

Definition 9.2. Let $T_{-1}M = \{v \in TM : v \text{ is future directed and } g(v,v) = -1\}$. Given $p \in M$, let $T_{-1}M|_p$ denote the fiber of $T_{-1}M$ at p. Also given $v \in T_{-1}M$, let c_v denote the unique timelike geodesic with $c_v'(0) = v$.

It is immediate from the reverse triangle inequality that if $\gamma : [0,a) \to M$ is a future directed nonspacelike geodesic and $d(\gamma(0), \gamma(s)) = L(\gamma|\,[0,s])$, then $d(\gamma(0), \gamma(t)) = L(\gamma|\,[0,t])$ for all t with $0 \leq t \leq s$. Also, the reverse triangle inequality implies that if $d(\gamma(0), \gamma(s)) > L(\gamma|\,[0,s])$, then $d(\gamma(0), \gamma(t)) > L(\gamma|\,[0,t])$ for all t with $s \leq t < a$. Hence the following definition makes sense.

Definition 9.3. Define the function $s : T_{-1}M \to \mathbb{R} \cup \{\infty\}$ by $s(v) = \sup\{t \geq 0 : d(\pi(v), c_v(t)) = t\}$.

We may first note that if $d(p,p) = \infty$, then $s(v) = 0$ for all $v \in T_{-1}M$ with $\pi(v) = p$. Also $s(v) > 0$ for all $v \in T_{-1}M$ if (M,g) is strongly causal. The number $s(v)$ may be interpreted as the "largest" parameter value t such that c_v is a maximal geodesic between $c_v(0)$ and $c_v(t)$. Indeed from Lemma 9.1 we know that the following result holds.

Corollary 9.4. *For $0 < t < s(v)$, the geodesic $c_v : [0,t] \to M$ is the unique maximal timelike curve (up to reparametrization) from $c_v(0)$ to $c_v(t)$.*

The function s fails to be upper semicontinuous for arbitrary space–times as may easily be seen by deleting a point from Minkowski space. But for timelike geodesically complete space–times we have the following proposition.

Proposition 9.5. *Let $v \in T_{-1}M$ with $s(v) > 0$. Suppose either that $s(v) = +\infty$, or $s(v)$ is finite and $c_v(t) = \exp(tv)$ extends to $[0, s(v)]$. Then s is upper semicontinuous at $v \in T_{-1}M$. Especially, if (M,g) is timelike geodesically complete, then $s : T_{-1}M \to \mathbb{R}\cup\{\infty\}$ is everywhere upper semicontinuous.*

Proof. It suffices to show the following. Let $v_n \to v$ in $T_{-1}M$ with $\{s(v_n)\}$ converging in $\mathbb{R} \cup \{\infty\}$. Then $s(v) \geq \lim s(v_n)$. If $s(v) = \infty$, there is nothing to prove. Hence we assume that $s(v) < \lim s(v_n) = A$ and $s(v) < \infty$ and derive a contradiction.

We may choose $\delta > 0$ such that $s(v) + \delta < A$ is in the domain of c_v and also assume that $s(v_n) \geq s(v) + \delta = b$ for all n. Let $c_n = c_{v_n}$. Since

$b \leq s(v_n)$, we have $d(\pi(v_n), c_n(b)) = b$ for all n. Since $v_n \to v$, we have by lower semicontinuity of distance that $d(\pi(v), c_v(b)) \leq \liminf d(\pi(v_n), c_n(b)) = b$.

Thus $d(\pi(v), c_v(b)) \leq b = L(c_v \mid [0, b])$, this last equality by definition of arc length. On the other hand, $d(\pi(v), c_v(b)) \geq L(c_v \mid [0, b])$ so that $d(\pi(v), c_v(b)) = L(c_v \mid [0, b]) = b$. Hence $s(v) \geq b = s(v) + \delta$, in contradiction. □

In order to prove the lower semicontinuity of s for globally hyperbolic space–times, it will first be useful to establish the following lemma.

Lemma 9.6. *Let (M, g) be a globally hyperbolic space–time, and let $\{p_n\}$ and $\{q_n\}$ be two infinite sequences of points with $p_n \to p$ and $q_n \to q$ where $p \ll q$. Assume $c_n : [0, d(p_n, q_n)] \to M$ is a unit speed maximal geodesic segment from p_n to q_n, and set $v_n = c_n{}'(0) \in T_{-1}M$. Then the sequence $\{v_n\}$ has a timelike limit vector $w \in T_{-1}M$. Moreover, $c_w : [0, d(p, q)] \to M$ is a maximal geodesic segment from p to q.*

Proof. By Corollary 3.32, there is a nonspacelike future directed limit curve c of c_n from p to q. By Proposition 3.34 and Remark 3.35, we have $L(c) \geq \limsup L(c_n) = \lim d(p_n, q_n) = d(p, q) > 0$. Thus as $d(p, q) \geq L(c)$, it follows that $L(c) = d(p, q) > 0$. Hence Theorem 4.13 implies that the curve c may be reparametrized to be a maximal timelike geodesic segment from p to q. Finally, $w = c'(0) \,/\, [-g(c'(0), c'(0))]^{1/2}$ is the required tangent vector. □

We are now ready to prove the lower semicontinuity of s for globally hyperbolic space–times.

Proposition 9.7. *If (M, g) is globally hyperbolic, then the function $s : T_{-1}M \to \mathbb{R} \cup \{\infty\}$ is lower semicontinuous.*

Proof. It suffices to prove that if $v_n \to v$ in $T_{-1}M$ and $s(v_n) \to A$ in $\mathbb{R} \cup \{\infty\}$, then $s(v) \leq A$. If $A = \infty$, there is nothing to prove. Thus suppose $A < \infty$. Assuming $s(v) > A$, we will derive a contradiction.

Choose $\delta > 0$ such that $A + \delta < s(v)$. (Since $A + \delta < s(v)$, the point $c_v(A + \delta)$ exists even if c_v does not extend to $s(v)$.) Define $b_n = s(v_n) + \delta$, and let N_0 be such that $b_n < s(v)$ for all $n \geq N_0$. Put $c_n = c_{v_n}$, $p_n = c_n(0)$, and $q = c_v(A+\delta)$. Since $v_n \to v$ and c_v is defined for some parameter values beyond $A+\delta$, the geodesics c_n must be defined for some parameter values past b_n whenever n is larger than some $N \geq N_0$. Let $q_n = c_n(b_n)$ for $n \geq N$. Now $c_n \mid [0, b_n]$ cannot be maximal since $b_n > s(v_n)$. Because M is globally hyperbolic and $c_n(0) \ll c_n(b_n)$, we may find maximal unit speed timelike geodesic segments $\gamma_n : [0, d(p_n, q_n)] \to M$ from p_n to q_n. Set $w_n = \gamma_n{}'(0)$. Since $c_v \mid [0, s(v))$ is a maximal geodesic and thus has no conjugate points, it is impossible for v to be a limit direction of $\{w_n : n \geq N\}$. Thus the maximal geodesic c_w joining p to q given by Lemma 9.6 applied to $\{w_n\}$ is different from c_v. This then implies that $s(v) \leq A+\delta$, which contradicts $A+\delta < s(v)$. □

The following example of a strongly causal space–time which is not globally hyperbolic shows that the hypothesis of global hyperbolicity is necessary for the lower semicontinuity of the function s in Proposition 9.7. Let $\mathbb{R}^2$ be given the usual Minkowski metric $ds^2 = dx^2 - dy^2$, and let (M, g) be the space–time formed by equipping $M = \mathbb{R}^2 - \{(1, y) \in \mathbb{R}^2 : 1 \leq y \leq 2\}$ with the induced metric (cf. Figure 9.1). Let $p = (0,0)$, $p_n = (1/n, 0)$, $v = \partial/\partial y|_p$, and $v_n = \partial/\partial y|_{p_n}$ for all $n \geq 1$. Then $v_n \to v$ as $n \to \infty$. Also let $\gamma(t) = (0, t)$ and $\gamma_n(t) = (p_n, t)$ for all $t \geq 0$. Conformally changing g on the compact set C shown in Figure 9.1 which is blocked from $I^+(p)$ by the slit $\{(1, y) \in \mathbb{R}^2 : 1 \leq y \leq 2\}$, we obtain a metric $\tilde{g}$ for M with the following properties. First γ is still a maximal geodesic in $(M, \tilde{g})$ so that $s(v) = +\infty$. But for each n there exists a timelike curve σ_n passing through the set C and joining p_n to a point $q_n = (1/n, y_n)$ on γ_n with $y_n \leq 4$ such that $L(\sigma_n) > L(\gamma_n[p_n, q_n])$. Hence $\gamma_n[p_n, q_n]$ fails to be maximal for all n so that $s(v_n) \leq 4$ for all n. Thus s is not lower semicontinuous. Note that $(M, \tilde{g})$ is strongly causal but $(M, \tilde{g})$ is not globally hyperbolic since $J^+((1,0)) \cap J^-((1,3))$ is not compact. The analogous construction may be applied to n-dimensional Minkowski space to produce a strongly causal n-dimensional space–time for which the function s fails to be lower semicontinuous.

Globally hyperbolic examples also may be constructed for which the function s is not upper semicontinuous. However, this discontinuity may occur at a tangent vector $v \in T_{-1}M$ in a globally hyperbolic space–time only if $s(v)$ is finite and $c_v(t)$ does not extend to $t = s(v)$. Combining Propositions 9.5 and 9.7, we obtain the following result.

Theorem 9.8. *Let (M, g) be globally hyperbolic, and suppose for $v \in T_{-1}M$ that either $s(v) = +\infty$, or $s(v)$ is finite and $c_v(t) = \exp(tv)$ extends to $[0, s(v)]$. Then s is continuous at $v \in T_{-1}M$. Especially, if (M, g) is globally hyperbolic and timelike geodesically complete, then $s : T_{-1}M \to \mathbb{R} \cup \{\infty\}$ is everywhere continuous.*

We are now ready to define the timelike cut locus.

Definition 9.9. (*Future and Past Timelike Cut Loci*) The *future timelike cut locus* $\Gamma^+(p)$ *in* T_pM is defined to be $\Gamma^+(p) = \{s(v)v : v \in T_{-1}M|_p$ and $0 < s(v) < \infty\}$. The *future timelike cut locus* $C_t^+(p)$ *of* p *in* M is defined to be $C_t^+(p) = \exp_p(\Gamma^+(p))$. If $0 < s(v) < \infty$ and $c_v(s(v))$ exists, then the point $c_v(s(v))$ is called the *future cut point* of $p = c_v(0)$ *along* c_v. The *past timelike cut locus* $C_t^-(p)$ and *past cut points* may be defined dually.

We may then interpret $s(v)$ as measuring the distance from p up to the future cut point along c_v. Thus Theorem 9.8 implies that for globally hyperbolic, timelike geodesically complete space–times, the distance from a fixed $p \in M$ to its future cut point in the direction $v \in T_{-1}M|_p$ is a continuous function of v.

We now give Lorentzian analogues of two well-known results relating cut and conjugate points on complete Riemannian manifolds. The following property of conjugate points is well known [Hawking and Ellis (1973, pp. 111–116), Lerner (1972, Theorem 4(6))].

Theorem 9.10. *A timelike geodesic is not maximal beyond the first conjugate point.*

In the language of Definition 9.9 this may be restated as follows.

Corollary 9.11. *The future cut point of $p = c_v(0)$ along c_v comes no later than the first future conjugate point of p along c_v.*

Utilizing this fact, we may now prove the second basic result on cut and conjugate points in globally hyperbolic space–times.

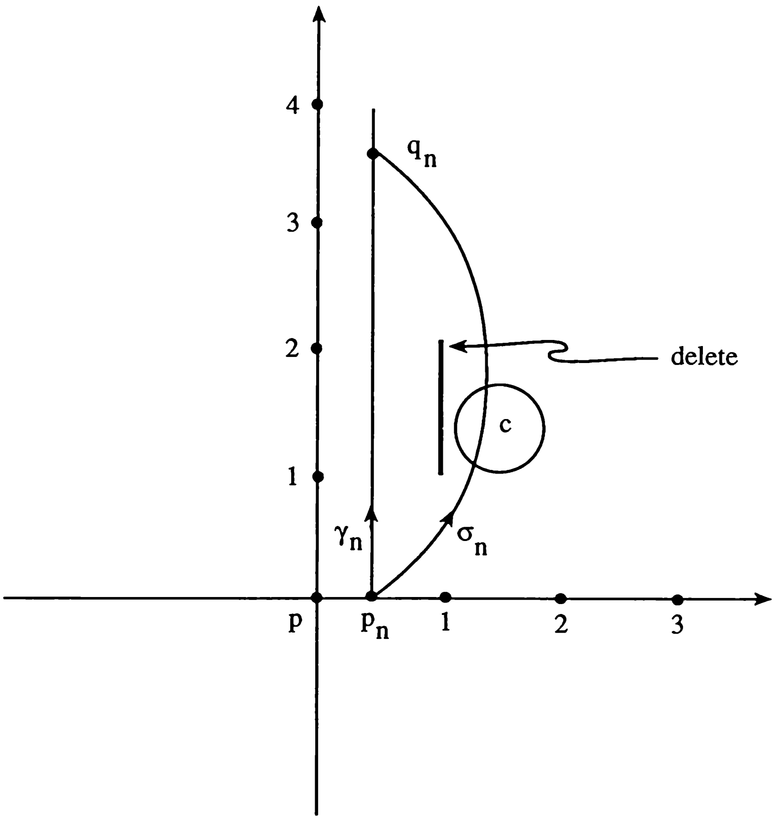


FIGURE 9.1. A strongly causal space–time $(M, \tilde{g})$ is shown with unit timelike tangent vectors v_n which converge to v, but with $s(v) = +\infty$ and $s(v_n) \leq 4$ for all $n \geq 1$. Hence $s : T_{-1}M \to \mathbb{R} \cup \{\infty\}$ is not lower semicontinuous.

Theorem 9.12. *Let (M, g) be globally hyperbolic. If $q = c(t)$ is the future cut point of $p = c(0)$ along the timelike geodesic c from p to q, then either one or possibly both of the following hold:*

(1) *The point q is the first future conjugate point of p along c.*

(2) *There exist at least two future directed maximal timelike geodesic segments from p to q.*

Proof. Without loss of generality we may suppose that $c = c_v$ for some $v \in T_{-1}M$ and thus that $t = d(p,q) = s(v)$. Let $\{t_n\}$ be a monotone decreasing sequence of real numbers converging to t. Since $c(t) \in M$, the points $c(t_n)$ exist for n sufficiently large. By global hyperbolicity, we may join $c(0)$ to $c(t_n)$ by a maximal timelike geodesic $c_n = c_{v_n}$ with $v_n \in T_{-1}M|_p$. Since $t_n > t = s(v)$, we have $v \neq v_n$ for all n. Let $w \in T_{-1}M$ be the timelike limiting vector for $\{v_n\}$ given by Lemma 9.6. If $v \neq w$, then c and c_w are two maximal timelike geodesic segments from p to q.

It remains to show that if $v = w$, then q is the first future conjugate point of p along c. If $v = w$, then there is a subsequence $\{v_m\}$ of $\{v_n\}$ with $v_m \to v$. If v were not a conjugate point, there would be a neighborhood U of v in $T_{-1}M|_p$ such that $\exp_p : U \to M$ is injective. On the other hand, since c_n and $c|\,[0,t_n]$ join $c(0)$ to $c(t_n)$ and $v_m \to v$, no such neighborhood U can exist. Thus q is a future conjugate point of p along c. By Corollary 9.11, q must be the first future conjugate point of p along c. □

Theorem 9.12 has the immediate implication that for globally hyperbolic space–times, $q \in C_t^+(p)$ if and only if $p \in C_t^-(q)$.

The timelike cut locus of a timelike geodesically complete, globally hyperbolic space–time has the following structural property which refines Theorem 9.12. We know from this theorem that if $q \in C_t^+(p)$ and q is not conjugate to p, then there exist at least two maximal geodesic segments from p to q. Accordingly, it makes sense to consider the set

$$\mathrm{Seg}(p) = \{q \in C_t^+(p) : \text{ there exist at least two future directed maximal geodesic segments from } p \text{ to } q\}.$$

Since $s : T_{-1}M \to \mathbb{R} \cup \{\infty\}$ is continuous by Theorem 9.8 and maximal geodesics joining any pair of causally related points exist in globally hyperbolic space–times, it may be shown that $\mathrm{Seg}(p)$ is dense in $C_t^+(p)$ for all $p \in M$. The proof may be given along the lines of Wolter's proof of Lemma 2 for complete

Riemannian manifolds [cf. Wolter (1979, p. 93)]. The dual result also holds for the past timelike cut locus $C_t^-(p)$.

9.2 The Null Cut Locus

The definition of null cut point has been given in Beem and Ehrlich (1979a, Section 5) where this concept was used to prove null geodesic incompleteness for certain classes of space–times. Let $\gamma : [0, a) \to (M, g)$ be a future directed null geodesic with endpoint $p = \gamma(0)$. Set $t_0 = \sup\{t \in [0, a) : d(p, \gamma(t)) = 0\}$. If $0 < t_0 < a$, we will say $\gamma(t_0)$ is the *future null cut point* of p on γ. Past null cut points are defined dually. Let $C_N^+(p)$ [respectively, $C_N^-(p)$] denote the *future* [respectively, *past*] *null cut locus* of p consisting of all future [respectively, past] null cut points of p. The definition of $C_N^+(p)$ together with the lower semicontinuity of distance yields $d(p, q) = 0$ for all $q \in C_N^+(p)$. We then define the *future nonspacelike cut locus* to be $C^+(p) = C_t^+(p) \cup C_N^+(p)$. The past nonspacelike cut locus is defined dually. For a subclass of globally hyperbolic space–times, Budic and Sachs (1976) have given a different but equivalent definition of null cut point using null generators for the boundary of $I^+(p)$.

The geometric significance of null cut points is similar to that of timelike cut points. The geodesic γ is maximizing from p up to and including the cut point $\gamma(t_0)$. That is, $L(\gamma|\,[0, t]) = d(p, \gamma(t)) = 0$ for all $t \leq t_0$. Thus there is no timelike curve joining p to $\gamma(t)$ for any t with $t \leq t_0$. In contrast, the geodesic γ is no longer maximizing beyond the cut point $\gamma(t_0)$. In fact, each point $\gamma(t)$ for $t_0 < t < a$ may be joined to p by a timelike curve.

Utilizing Proposition 2.19 of Penrose (1972, p. 15) and the definition of maximality, the following lemma is easily established.

Lemma 9.13. *Let (M, g) be a causal space–time. If there are two future directed null geodesic segments from p to q, then q comes on or after the null cut point of p on each of the two segments.*

The cylinder $S^1 \times \mathbb{R}$ with Lorentzian metric $ds^2 = d\theta dt$ shows that the assumption that (M, g) is causal is needed in Lemma 9.13.

We next prove the null analogue of Lemma 9.6.

Lemma 9.14. *Let (M, g) be globally hyperbolic, and let p, q be distinct points in M with $p \leqslant q$ and $d(p, q) = 0$. Assume that $p_n \to p$, $q_n \to q$ and $p_n \leqslant q_n$. Let c_n be a maximal geodesic joining p_n to q_n with initial direction v_n. Then the set of direction $\{v_n\}$ has a limit direction w, and c_w is a maximal null geodesic from p to q.*

Proof. Using Corollary 3.32 we obtain a future directed nonspacelike limit curve λ from p to q. Since $d(p, q) = 0$, the curve λ must satisfy $L(\lambda) = 0$. It follows that λ may be reparametrized to a maximal null geodesic. □

We may now obtain the null analogue of Theorem 9.12.

Theorem 9.15. *Let (M, g) be globally hyperbolic and let $q = c(t)$ be the future null cut point of $p = c(0)$ along the null geodesic c. Then either one or possibly both of the following hold:*

(1) *The point q is the first future conjugate point of p along c.*
(2) *There exist at least two future directed maximal null geodesic segments from p to q.*

Proof. Let $v = c'(0)$, and let t_n be a monotone decreasing sequence of real numbers with $t_n \to t$. Since $q \in M$, we know that $c(t_n)$ exists for all sufficiently large n. Since (M, g) is globally hyperbolic, we may find maximal nonspacelike geodesics c_n with initial directions v_n joining p to $c(t_n)$. By Lemma 9.14 the set of directions $\{v_n\}$ has a limit direction w. If $v \neq w$, then the geodesic c_w is a second maximal null geodesic joining p to q as $d(p, q) = 0$. If $v = w$, then q is conjugate to p along c. Since $d(p, q) = 0$, q must be the first conjugate point of c (cf. Theorem 10.72). □

We now show how the null cut locus may be applied to prove theorems on the stability of null geodesic incompleteness. The key ideas needed for this application are first, that many physically interesting space–times may be conformally embedded in a portion of the Einstein static universe (cf. Example 5.11) that is free of null cut points and second, that null cut points are invariant under conformal changes of metric. A discussion of global conformal embeddings of anti–de Sitter space–time and the Friedmann cosmological models in the Einstein static universe is given in Penrose (1968) [cf. Hawking and Ellis (1973, pp. 132, 141)].

We now digress briefly to give an explicit computational discussion of the well-known fact that null pregeodesics are invariant under global conformal changes. An indirect proof of the conformal invariance of null cut points may also be given using the Lorentzian distance function.

Recall that a smooth curve $\gamma : J \to M$ is said to be a *pregeodesic* if γ can be reparametrized to a smooth curve c which satisfies the geodesic differential equation $\nabla_{c'}c'(t) = 0$. Also recall the following definition.

Definition 9.16. (*Global Conformal Diffeomorphism*) A diffeomorphism $f : (M_1, g_1) \to (M_2, g_2)$ of M_1 onto M_2 is said to be a *global conformal diffeomorphism* if there exists a smooth function $\Omega : M_1 \to \mathbb{R}$ such that

$$f^* g_2 = e^{2\Omega} g_1.$$

The space–time (M_1, g_1) is said to be *globally conformally diffeomorphic to an open subset* U of (M_2, g_2) if there exists a diffeomorphism $f : M_1 \to U$ and a smooth function $\Omega : M_1 \to \mathbb{R}$ such that

$$f^*(g_2|_U) = e^{2\Omega} g_1.$$

The purpose of using the factor $e^{2\Omega}$ rather than just a positive-valued smooth function in Definition 9.16 is to give the simplest possible formula relating the connections ∇ for (M_1, g_1) and $\widetilde{\nabla}$ for (M_2, g_2). Explicitly, it may be shown that if $f^* g_2 = e^{2\Omega} g_1$ where $f : (M_1, g_1) \to (M_2, g_2)$ is any smooth function, then

$$\text{(9.1)} \qquad \widetilde{\nabla}_{f_* X} f_* Y \Big|_{f(p)} = f_* \nabla_X Y\big|_p + X_p(\Omega) f_* Y(p) + Y_p(\Omega) f_* X(p) - g_1(X(p), Y(p)) f_*(\operatorname{grad} \Omega(p))$$

for any pair X, Y of vector fields on M_1. Here "grad Ω" denotes the gradient vector field of Ω with respect to the metric g_1 for M_1.

Using formula (9.1), it is now possible to show that if $\gamma : J \to M_1$ is a null geodesic in (M_1, g_1), then $\sigma = f \circ \gamma : J \to M_2$ is a null pregeodesic in (M_2, g_2).

The crux of the matter is that because $\gamma'(t)$ is null and $\nabla_{\gamma'}\gamma' = 0$, formula (9.1) simplifies to

$$\widetilde{\nabla}_{\sigma'}\sigma'(t) = 2\gamma'(t)(\Omega)\sigma'(t)$$

for all $t \in J$. Note, however, that if γ were a timelike geodesic, then the factor $g_1(\gamma', \gamma') f_*(\operatorname{grad}\Omega)$ in formula (9.1) would prevent $f \circ \gamma$ from being a timelike pregeodesic in (M_2, g_2).

Lemma 9.17. *Let $f : (M_1, g_1) \to (M_2, g_2)$ be a global conformal diffeomorphism of M_1 onto M_2. Then $\gamma : J \to (M_1, g_1)$ is a null pregeodesic for (M_1, g_1) iff $f \circ \gamma$ is a null pregeodesic for (M_2, g_2).*

Proof. Since $f^{-1} : (M_2, g_2) \to (M_1, g_1)$ is also a conformal diffeomorphism, it is enough to show that if $\gamma : J \to M_1$ is a null geodesic, then $\sigma = f \circ \gamma$ is a null pregeodesic of (M_2, g_2). That is, we must show that σ can be reparametrized to be a null geodesic of (M_2, g_2). But it has already been shown that

$$\widetilde{\nabla}_{\sigma'}\sigma'(t) = f(t)\sigma'(t) \tag{9.2}$$

for some smooth function $f : J \to \mathbb{R}$. Just as in the classical theory of projectively equivalent connections in Riemannian geometry, however, formula (9.2) implies that σ is a pregeodesic [cf. Spivak (1970, pp. 6–37 ff.)]. □

We are now ready to prove the conformal invariance of null cut points under global conformal diffeomorphisms $f : (M_1, g_1) \to (M_2, g_2)$. Notice that if (M_1, g_1) is time oriented by the vector field X_1, then (M_2, g_2) is time oriented either by $X_2 = f_* X_1$ or $-X_2$. If M_2 is time oriented by X_2, then f maps future directed curves to future directed curves and thus future (respectively, past) null cut points to future (respectively, past) null cut points in Proposition 9.18. On the other hand, if M_2 is time oriented by $-X_2$, then f maps future (respectively, past) null cut points to past (respectively, future) null cut points since f maps future directed curves to past directed curves.

Proposition 9.18. *Let $f : (M_1, g_1) \to (M_2, g_2)$ be a global conformal diffeomorphism of (M_1, g_1) onto (M_2, g_2). Let $\gamma : [0, a) \to (M_1, g_1)$ be a null geodesic of M_1. If $q = \gamma(t_0)$ is a null cut point of $p = \gamma(0)$ along the* *geodesic γ, then $f(q)$ is a null cut point of $f(p)$ along the null pregeodesic $f \circ \gamma : [0, a) \to (M_2, g_2)$.*

Proof. It suffices to assume that q is the future null cut point of p along γ and that (M_2, g_2) is time oriented so that $f \circ \gamma$ is a future directed null curve in M_2. Let $\sigma : [0, b) \to (M_2, g_2)$ be a reparametrization of $f \circ \gamma$ to a future directed null geodesic guaranteed by Lemma 9.17 with $f(p) = \sigma(0)$. Then $f(q) = \sigma(t_1)$ for some $t_1 \in (0, b)$. Let d_i denote the Lorentzian distance function of (M_i, g_i) for $i = 1, 2$.

We show first that $d_2(\sigma(0), \sigma(t)) = 0$ for any t with $0 \leq t \leq t_1$. For suppose $d_2(\sigma(0), \sigma(t)) \neq 0$ for some t with $0 \leq t \leq t_1$. Then we may find a future directed nonspacelike curve β in M_2 from $\sigma(0)$ to $\sigma(t)$ with $L_{g_2}(\beta) > 0$. Thus $f^{-1} \circ \beta$ is a future directed nonspacelike curve in M_1 from p to $f^{-1}(\sigma(t))$ with $L_{g_1}(f^{-1} \circ \beta) > 0$. Since $f(q) = \sigma(t_1)$, we have $f^{-1}(\sigma(t)) = \gamma(t_2)$ for some $t_2 \leq t_0$. Hence $d_1(p, \gamma(t_2)) \geq L_{g_1}(f^{-1} \circ \beta) > 0$, which contradicts the fact that $d_1(p, \gamma(t_2)) = 0$ since $t_2 \leq t_0$ and $\gamma(t_0) = q$ was the future null cut point to p along γ.

We show now that $d_2(\sigma(0), \sigma(t)) \neq 0$ for any $t > t_1$. This then makes $f(q) = \sigma(t_1)$ the future null cut point of $f(p)$ along σ as required. To this end, fix $t > t_1$. There is then a $t_2 > t_0$ so that $f^{-1}(\sigma(t)) = \gamma(t_2)$. Since $\gamma(t_0) = q$ is the future null cut of p along γ, we have $d_1(p, \gamma(t_2)) > 0$. Hence there is a future directed nonspacelike curve α from p to $\gamma(t_2)$ with $L_{g_1}(\alpha) > 0$. Then $f \circ \alpha$ is a future directed nonspacelike curve from $f(p)$ to $\sigma(t)$. Hence $d_2(f(p), \sigma(t)) \geq L_{g_2}(f \circ \alpha) > 0$ as required. □

With Proposition 9.18 in hand, we may now apply the null cut locus to study null geodesic incompleteness. Recall that a geodesic is said to be incomplete if it cannot be extended to all values of an affine parameter (cf. Definition 6.2).

Let R and Ric denote the curvature tensor and Ricci curvature tensor of (M, g), respectively. Recall that an inextendible null geodesic γ will satisfy the generic condition if for some parameter value t, there exists a nonzero tangent vector $v \in T_{\gamma(t)}M$ with $g(v, \gamma'(t)) = 0$ such that $R(v, \gamma'(t))\gamma'(t)$ is nonzero and is *not* proportional to $\gamma'(t)$ (cf. Proposition 2.11, Section 2.5). In particular, if $\mathrm{Ric}(v, v) > 0$ for all null vectors $v \in TM$, then every inextendible null geodesic of (M, g) satisfies the generic condition. In Section 2.5, it was noted that dim $M \geq 3$ is necessary for the null generic condition to be satisfied. Thus we will assume that dim $M \geq 3$ in the following proposition.

Proposition 9.19. *Let (M, g) be a space–time of dimension $n \geq 3$ such that all inextendible null geodesics satisfy the generic condition and such that* $\operatorname{Ric}(v, v) \geq 0$ *for all null vectors. If (M, g) is globally conformally diffeomorphic to an open subset of a space–time (M', g') which has no null cut points, then all null geodesics of (M, g) are incomplete.*

Proof. Proposition 4.4.5 of Hawking and Ellis (1973, p. 101) shows that if the Ricci curvature is nonnegative on all null vectors, then each complete null geodesic which satisfies the generic condition has conjugate points (cf. Proposition 12.17). Since maximal geodesics do not contain conjugate points, we need only show that all null geodesics of (M, g) are maximal to establish their incompleteness. Assume that γ is a future directed null geodesic from p to q which is not maximal. Then there is a timelike curve from p to q. Since conformal diffeomorphisms take null geodesics to null pregeodesics and timelike curves to timelike curves, the image of γ must be a nonmaximal null geodesic in (M', g'). This implies that (M', g') has a null cut point and hence yields a contradiction. □

Recall that the four-dimensional Einstein static universe (Example 5.11) is the cylinder $x^2+y^2+z^2+w^2 = 1$ embedded in $\mathbb{R}^5$ with the metric induced from the Minkowski metric $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2 + dw^2$. The Einstein static universe is thus $\mathbb{R} \times S^3$ with a Lorentzian product metric. The geodesics and null cut points are easy to determine in this space–time. The null cut locus of the point (t, x, y, z, w) merely consists of the two points $(t \pm \pi, -x, -y, -z, -w)$. Consequently, the subset

$$M' = \{(t, x, y, z, w) : 0 < t < \pi, \quad x^2 + y^2 + z^2 + w^2 = 1\}$$

has the property that $C_N(p) \cap M' = \emptyset$ for all $p \in M'$. Proposition 9.19 then has the following consequence.

Corollary 9.20. *Let (M, g) be a space–time such that all null geodesics satisfy the generic condition and such that* $\operatorname{Ric}(v, v) \geq 0$ *for all null vectors. If*

(M, g) is globally conformally diffeomorphic to some portion of the subset M' of the Einstein static universe, then all null geodesics of (M, g) are incomplete.

Since Minkowski space–time is free of null cut points, the above result remains valid if we replace M' with Minkowski space–time.

Friedmann space–times are used in cosmology as models of the universe. In these spaces it is assumed that the universe is filled with a perfect fluid having zero pressure. We will also assume that the cosmological constant Λ is zero for these models. These space–times are then Robertson–Walker spaces (cf. Section 5.4) with $p = \Lambda = 0$. These space–times may be conformally embedded in the subset M' of the Einstein static universe defined above [cf. Hawking and Ellis (1973, pp. 139–141)]. Furthermore, $\mathrm{Ric}(g)(v, v) > 0$ for all nonspacelike vectors in a Friedmann cosmological model (M, g). By Proposition 7.3, there is a C^2 neighborhood $U(g)$ of g in $C(M, g)$ such that $\mathrm{Ric}(g_1)(v, v) > 0$ for all $g_1 \in U(g)$ and all nonspacelike vectors v in (M, g_1). Since $\mathrm{Ric}(g_1)(v, v) > 0$ implies that the generic condition is satisfied by all null geodesics in (M, g_1), Corollary 9.20 yields the following result.

Corollary 9.21. *Let (M, g) be a Friedmann cosmological model. There is a C^2 neighborhood $U(g)$ of g in $C(M, g)$ such that every null geodesic in (M, g_1) is incomplete for all $g_1 \in U(g)$.*

9.3 The Nonspacelike Cut Locus

Recall the following definitions.

Definition 9.22. (*Future and Past Nonspacelike Cut Loci*) The *future nonspacelike cut locus* $C^+(p)$ *of* p is defined as $C^+(p) = C_t^+(p) \cup C_N^+(p)$. The *past nonspacelike cut locus* is $C^-(p) = C_t^-(p) \cup C_N^-(p)$, and the *nonspacelike cut locus* is $C(p) = C^-(p) \cup C^+(p)$.

We mentioned in Chapter 8 that a complete noncompact Riemannian manifold admits a geodesic ray issuing from each point. Thus at each point there is a direction such that the geodesic issuing from that point in that direction has no cut point. For strongly causal space–times, the Lorentzian analogue of this property follows similarly from the existence of past and future directed nonspacelike geodesic rays issuing from each point. This may be restated as the first half of the following proposition.

Proposition 9.23. *(1) Let (M, g) be strongly causal. Then given any point $p \in M$, there is a future and a past directed nonspacelike tangent vector in T_pM such that the geodesics issuing from p in these directions have no cut points.*

(2) Let (M, g) be globally hyperbolic. Given any point $p \in M$, p has no farthest nonspacelike cut point.

Proof. As we have already remarked, part (1) follows immediately from Theorem 8.10. For the proof of part (2), assume $q \in M$ is a farthest cut point of p. Then q is a cut point along a maximal geodesic segment γ from p to q. Choose a sequence of points $\{q_n\}$ such that $q \ll q_n$ for each n and $q_n \to q$. Since (M, g) is globally hyperbolic, there exist maximal timelike geodesic segments $c_n : [0, d(p, q_n)] \to M$ from p to q_n for each n. Extend each c_n to a future inextendible geodesic. Since q is a farthest cut point of p and $d(p, q_n) \geq d(p, q) + d(q, q_n) > d(p, q)$, for each n the geodesic ray c_n contains no cut point of p. The sequence $\{c_n\}$ has a limit curve c which is a future directed and future inextendible nonspacelike curve starting at p by Proposition 3.31. By passing to a subsequence if necessary, we may assume that $\{c_n\}$ converges to c in the C^0 topology on curves (cf. Proposition 3.34). Using the global hyperbolicity of (M, g) and the fact that $q_n \to q$, we find that $q \in c$. If $r \in c$ and $r_n \in c_n$ with $r_n \to r$, then $d(p, r) = \lim d(p, r_n)$. Using the upper semicontinuity of arc length for strongly causal space–times, we find that the length of c from p to r is at least as great as $\limsup d(p, r_n) = \lim d(p, r_n) = d(p, r)$. Thus c is a maximal geodesic ray. Since q is a cut point of p on γ, the geodesics c and γ are distinct maximal nonspacelike geodesics containing p and q. Either Lemma 9.1 or Lemma 9.13 now yields a contradiction to the maximality of c beyond q. $\square$

We now turn to the analogue of Klingenberg's observation that, for Riemannian manifolds, if q is a closest cut point to p and q is nonconjugate to p, then there is a geodesic loop at p passing through q. For Lorentzian manifolds, however, a different result is true. If $\{q_n\} \subseteq C_t^+(p)$ converges to $q \in C_N^+(p)$, then $d(p, q_n) \to 0$ in a globally hyperbolic space–time since the Lorentzian distance function is continuous. Thus it is not unreasonable to expect that, for globally hyperbolic space–times, there is no closest timelike cut point q to p that is nonconjugate to p.

Theorem 9.24. *Let (M, g) be a globally hyperbolic space–time, and assume that $p \in M$ has a closest future (or past) nonspacelike cut point q. Then q is either conjugate to p, or else q is a null cut point of p.*

Proof. Let q be a future cut point of p which is a closest cut point of p with respect to the Lorentzian distance d. Assume q is neither conjugate to p nor a null cut point of p. Then $p \ll q$, and by Theorem 9.12 there exist at least two future directed maximal timelike geodesics c_1 and c_2 from p to q. Let $\gamma : [0, a) \to M$ be a past directed timelike curve starting at q. By choosing $a > 0$ sufficiently small, we may assume the image of γ lies in the chronological future of p. Then $p \ll \gamma(t) \ll q$ for $0 < t < a$ implies $d(p, q) \geq d(p, \gamma(t)) + d(\gamma(t), q) > d(p, \gamma(t))$ using the reverse triangle inequality. Since q is a closest cut point, the point $\gamma(t)$ comes before a cut point of p on any timelike geodesic from p to $\gamma(t)$. Thus any timelike geodesic from p to $\gamma(t)$ is maximal. Since q is not conjugate to p along c_1, there is a timelike geodesic from p to $\gamma(t)$ near c_1 for all sufficiently small t. Similarly, there exists a timelike geodesic from p to $\gamma(t)$ near c_2 for all small t. The existence of two maximal timelike geodesics from p to $\gamma(t)$ implies $\gamma(t)$ is a cut point of p and yields a contradiction since $d(p, \gamma(t)) < d(p, q)$. □

For Riemannian manifolds, the set of unit tangent vectors in any given tangent space is compact. It is then an immediate consequence of the Riemannian analogues of Propositions 9.5 and 9.7 above that the cut locus of any point in a complete Riemannian manifold is a closed subset of M [cf. Gromoll, Klingenberg, and Meyer (1975, pp. 170–171), Kobayashi (1967, pp. 100–101)]. For Lorentzian manifolds, the timelike cut locus is not a closed subset of M in general as the Einstein static universe (Example 5.11) shows. However, for globally hyperbolic space–times it may be shown that the nonspacelike cut locus and the null cut locus of any point are closed subsets of M [cf. Beem and Ehrlich (1979c, Proposition 6.5)]. It may be seen by Example 9.28 given below that the assumption of global hyperbolicity is necessary for the nonspacelike cut locus to be a closed subset of M. An "intrinsic" proof for Lorentzian manifolds is more complicated than the proof for Riemannian manifolds as a result of the lack of unit null tangent vectors.

We now turn to the proof of the closure of the nonspacelike and null cut loci for globally hyperbolic space–times. We first establish a technical lemma. Here the assumption that $\exp_p(v) = q$ is essential to the validity of the conclusion of this lemma.

Lemma 9.25. *Let (M, g) be a strongly causal space–time. Assume $p \leqslant q_n$ and $q_n \to q \neq p$. If $\exp_p(v_n) = q_n$, $\exp_p(v) = q$, and the directions of the nonspacelike vectors v_n converge to the direction of v, then $v_n \to v$.*

Proof. Choose numbers $a_n > 0$ such that the vectors $w_n = a_n v_n$ converge to v. Then the sequence of points $r_n = \exp_p(w_n)$ must be defined for large n and must converge to q. Since (M, g) is causal, there is only one value t_n of t such that $\exp_p(t_n w_n) = q_n$, namely $t_n = a_n^{-1}$. Using $r_n \to q$, $q_n \to q$, and the strong causality of (M, g) near q, we find that $t_n \to 1$. Since $v_n = t_n w_n$, the sequence $\{v_n\}$ converges to v. □

Fix a point $p \in M$. A tangent vector $v \in T_pM$ is a conjugate point to p in T_pM if $(\exp_p)_*$ is singular at v. The conjugate points to p in T_pM must form a closed subset of the domain of $\exp_p$ because $(\exp_p)_*$ is nonsingular on an open subset of T_pM. A point $q \in M$ is conjugate to p along a geodesic c if there is some conjugate point $v \in T_pM$ such that $\exp_p v = q$ and c is (up to reparametrization) the geodesic $\exp_p(tv)$. If $q \in M$, $p \leqslant q$, and q is conjugate to p along a nonspacelike geodesic, then we will say q is a *future nonspacelike conjugate point of p*. Past nonspacelike conjugate points are defined dually. If (M, g) is a causal space–time, then p cannot be in its own future or past nonspacelike conjugate locus because a nonspacelike geodesic passes through p at most once and $(\exp_p)_*$ is nonsingular at the origin of T_pM.

Remark 9.26. It is possible even in globally hyperbolic space–times for a point to be conjugate to itself along *spacelike* geodesics, however. This happens in any Einstein static universe of dimension $n \geq 3$.

Lemma 9.27. *Let (M, g) be a globally hyperbolic space–time. Then the future (respectively, past) nonspacelike conjugate locus in M is a closed subset of M.*

Proof. Assume that $\{q_n\}$ is a sequence of future nonspacelike conjugate points of p with $q_n \to q$. Then $p \neq q$, $p \leqslant q$, and $p \leqslant q_n$ for each n. Let $v_n \in T_pM$ be a nonspacelike vector such that $(\exp_p)_*$ is singular at v_n and $\exp_p(v_n) = q_n$. Then $c_n(t) = \exp_p(tv_n)$ is a future directed, future inextendible geodesic starting at p and containing q_n. The geodesics c_n must have a limit curve γ by Proposition 3.31. Since (M, g) is globally hyperbolic, γ must contain q. Using the strong causality of (M, g) and the fact that γ is a limit curve of nonspacelike geodesics, we find that γ is itself a nonspacelike geodesic. Let v be the unique (nonspacelike) vector tangent to γ at $\gamma(0) = p$ such that $\exp_p(v) = q$. Since γ is a limit curve of $\{c_n\}$, there must be some subsequence $\{m\}$ of $\{n\}$ such that the directions of the vectors v_m converge to the direction of v. Lemma 9.25 then implies that $v_m \to v$. Since the vectors v_m belong to the total conjugate locus of p in T_pM and this conjugate locus is a closed subset of the domain of $\exp_p$, the vector v is a conjugate point of p in T_pM. Because v is a future directed nonspacelike vector, the point $q = \exp_p(v)$ is a future nonspacelike conjugate point of p in M. This establishes the closure in M of the future nonspacelike conjugate locus. The dual proof shows that the past nonspacelike conjugate locus is also closed. □

At this juncture we stress that Lemma 9.27 reflects a specifically non-Riemannian phenomenon. Contrary to well-established folk lore, the (first) conjugate locus, even for a compact Riemannian manifold, need *not* be closed in general [cf. Magerin (1993)]. In the space–time case we are rescued by the non-imprisonment property that a nonspacelike geodesic must escape from any compact subset of a globally hyperbolic space–time in finite affine parameter. On the other hand, examples may be constructed of stably causal space–times which do not have closed nonspacelike conjugate loci by deleting appropriate subsets from the four-dimensional Einstein static universe.

Example 9.28. Let $M = S^1 \times \mathbb{R} = \{(\theta, t) : 0 \leq \theta \leq 2\pi,\ t \in \mathbb{R}\}$ with the flat metric $ds^2 = d\theta^2 - dt^2$. This is the two-dimensional Einstein static universe which is globally hyperbolic. There are no conjugate points in this space–time. If $p = (0,0)$, the future null cut locus $C_N^+(p)$ consists of the single point (π, π). The future timelike cut locus $C_t^+(p)$ consists of $\{(\pi, t) : t > \pi\}$ and is not closed. On the other hand, $C^+(p) = C_t^+(p) \cup C_N^+(p)$ is closed.

If we delete the two points $(\pm\pi/4, 2\pi)$ from M, we obtain a strongly causal space–time $(M', g|_{M'})$ which is not globally hyperbolic. In $(M', g|_{M'})$ the future nonspacelike cut locus $C^+(p)$ is *not* closed.

Proposition 9.29. *Let (M,g) be a globally hyperbolic space–time. For any $p \in M$, the sets $C_N^+(p)$, $C_N^-(p)$, $C^+(p)$, and $C^-(p)$ are all closed subsets of M. In particular, the null cut locus and the nonspacelike cut locus of p are each closed.*

Proof. Since the four cases are similar, we will show only that $C_N^+(p)$ is closed in M. To this end, let $q_n \in C_N^+(p)$ and $q_n \to q$. Since (M,g) is globally hyperbolic, $q \neq p$ and $q \in J^+(p)$. From the definition of $C_N^+(p)$ we see that $d(p, q_n) = 0$. Using the continuity of Lorentzian distance for globally hyperbolic space–times (Lemma 4.5), it follows that $d(p,q) = \lim d(p, q_n) = 0$. Let γ be any nonspacelike geodesic from p to q. Then $d(p,q) = 0$ implies that γ is a maximal null geodesic and that the cut point to p along γ cannot come before q.

There are two cases to consider. Either infinitely many points q_n are future nonspacelike conjugate to p or else no q_n is future nonspacelike conjugate to p for large n. In the first case we apply Lemma 9.27 and find that if q is future nonspacelike conjugate to p, then if this conjugacy is along γ, q is the cut point along γ because the cut point along γ comes before or at the first conjugate point along γ. If q is a future nonspacelike conjugate point to p along some other nonspacelike geodesic γ', then γ' must be null, and Lemma 9.13 shows that q is the cut point along both γ and γ'.

Assume now that q_n is not future nonspacelike conjugate to p for all sufficiently large n. Theorem 9.15 implies that for large n there exist at least two maximal null geodesics from p to q_n. Thus for large n there are two future directed null vectors v_n and w_n with $\exp_p(v_n) = \exp_p(w_n) = q_n$. The future directed nonspacelike directions at p form a compact set of directions. Thus the directions determined by $\{v_n\}$ and $\{w_n\}$ have limit directions v and w respectively. If v and w determine different directions, we apply Lemma 9.14 to obtain two maximal null geodesics from p to q. In this case q is a cut point of p. On the other hand, v and w may determine the same direction. If this is so, we first note that there are constants $a > 0$ and $b > 0$ such that $av = bw$ and $\exp_p(av) = \exp_p(bw) = q$. Applying Lemma 9.25, it follows that for some subsequence $\{m\}$ of $\{n\}$ we have $v_m \to av$ and $w_m \to av$. However $v_m \neq w_m$ and $\exp_p(v_m) = \exp_p(w_m)$ then yield that $(\exp_p)_*$ must be singular at av. Thus q is conjugate to p along $\gamma(t) = \exp_p(tav)$, and since $d(p, q) = 0$ we find that q is the cut point of p along $\gamma(t)$. □

Surprisingly, the above result holds without any assumptions about the timelike or null geodesic completeness of (M, g). In particular, the null cut locus and the nonspacelike cut locus in globally hyperbolic space–times are closed by Proposition 9.29, even though the function $s : T_{-1}M \to \mathbb{R} \cup \{\infty\}$ may fail to be upper semicontinuous.

It is well known [cf. Kobayashi (1967, pp. 100–101)] that using the cut locus, it may be shown that a compact Riemannian manifold is the disjoint union of an open cell and a closed subset (the cut locus of a fixed $p \in M$) which is a continuous image (under $\exp_p$) of an $(n-1)$-sphere. Thus the cut loci inherit many of the topological properties of the compact manifold itself. For Lorentzian manifolds, cut points may be defined (using the Lorentzian distance at least) only for nonspacelike geodesics. Hence a corresponding result for space–times must describe the topology not of all of M itself, but only of $J^+(p)$ for an arbitrary $p \in M$. To obtain this decomposition with the methodology of this section, we need to assume that (M, g) is timelike geodesically complete as well as globally hyperbolic, so that the function $s : T_{-1}M \to \mathbb{R} \cup \{\infty\}$ defined in Definition 9.3 will be continuous and not just lower semicontinuous (cf. Propositions 9.5 and 9.7).

Recall also that the future horismos $E^+(p)$ of any point $p \in M$ is given by $E^+(p) = J^+(p) - I^+(p)$ and that $C^+(p)$ denotes the future nonspacelike cut locus of p.

Proposition 9.30. *Let (M, g) be globally hyperbolic and timelike geodesically complete. For each $p \in M$ the set $J^+(p) - [C^+(p) \cup E^+(p)]$ is an open cell.*

Proof. Let $B = J^+(p) - [C^+(p) \cup E^+(p)]$. Then $q \in B$ if and only if there is a maximal future directed timelike geodesic which starts at p and extends beyond q. Thus $B = I^+(p) - C^+(p)$ which shows that B is open. Now $T_{-1}M|_p = \{v \in T_pM : v \text{ is future directed and } g(v,v) = -1\}$ is homeomorphic to $\mathbb{R}^{n-1}$. Let $H : T_{-1}M|_p \to \mathbb{R}^{n-1}$ be a homeomorphism, and define $\bar{s} : \mathbb{R}^{n-1} \to \mathbb{R} \cup \{\infty\}$ by $\bar{s} = s \circ H^{-1}$. There is an induced homeomorphism of B with $\{(x,t) \in \mathbb{R}^{n-1} \times \mathbb{R} : 0 < t < \bar{s}(x)\}$ defined by $q \to (H(v), t)$, where v is the vector in $T_{-1}M$ such that $\exp_p(sv)$ is the unique (up to reparametrization) maximal geodesic from p to q, and $\exp_p(tv) = q$ with $v \in T_{-1}M$. Let $f : [0, \infty] \to [0, 1]$ be a homeomorphism with $f(0) = 0$ and $f(\infty) = 1$. Then the map $(x,t) \to (x, f(t) \,/\, f(\bar{s}(x)))$ shows B is homeomorphic to $\mathbb{R}^{n-1} \times (0,1)$, which establishes the proposition. □

Since $I^+(p) \subseteq J^+(p)$, Proposition 9.30 yields some indirect topological information about $I^+(p)$. The Einstein static universe shows that $I^+(p)$ need not be an open cell even if (M, g) is globally hyperbolic. However, $(I^+(p), g|_{I^+(p)})$ is globally hyperbolic whenever (M, g) is globally hyperbolic. Thus $I^+(p)$ may be expressed topologically as a product $I^+(p) = S \times \mathbb{R}$, where S is an $(n-1)$-dimensional manifold (cf. Theorem 3.17).

9.4 The Nonspacelike Cut Locus Revisited

In Sections 9.2 and 9.3 we treated the null and timelike cut loci separately because the intrinsic unit observer bundle $T_{-1}M$ was used as the primary technical tool for investigating the timelike cut locus. This approach suffers the drawback, however, that if $\{q_n\}$ is a sequence of future timelike cut points to p and $q_n = \exp(s(v_n)v_n)$ converges to a null cut point q to p, then

$$\lim s(v_n) = 0. \tag{9.3}$$

As a result of the technical differences in handling the null and timelike cut loci, the proof in Proposition 9.29 of the closure of the nonspacelike and null cut loci for globally hyperbolic space–times could not draw on the continuity properties of the s-function established in Proposition 9.7. Rather, we relied on a more indirect proof method which called upon the characterization of nonspacelike cut points given in Theorems 9.12 and 9.15. In this section we will give a more elementary method to derive the closure of the nonspacelike cut loci for globally hyperbolic space–times which treats both null and timelike cut points simultaneously but is non-intrinsic. The treatment given in this section has been worked out in discussions with G. Galloway.

We have noted in Section 9.3 just prior to Example 9.28 that Magerin (1993) has given examples for Riemannian spaces that show that, contrary to folk lore, the first conjugate locus of a point in a compact Riemannian manifold need not be closed. Thus for completeness, we present an elementary proof that the cut locus of any point in a complete Riemannian manifold is closed, nonetheless.

Theorem 9.31. *Let (N, g_0) be a complete Riemannian manifold, and let p in N be arbitrary. Then the cut locus $C(p)$ of p in N is a closed subset of N.*

Proof. We have the basic fact that the distance to the cut point in unit direction v, which we will denote as is customary by $s : SN \to \mathbb{R} \cup \{\infty\}$ corresponding to Definition 9.3 for unit timelike vectors, is continuous provided (N, g_0) is complete. Now let $\{q_n\} \subseteq C(p)$ be given with $q_n \to q$ in M. Represent $q_n = \exp_p(s(v_n)v_n)$ with $\|v_n\| = 1$. Since the set of unit vectors in T_pM is compact, we may assume by passing to a subsequence if necessary that $v_n \to v \in T_pM$. By continuity of the s-function, $\lim s(v_n)v_n = s(v)v$. Since (N, g_0) is geodesically complete, $\exp_p(s(v)v)$ is defined. Hence

$$q = \lim q_n = \lim \exp_p(s(v_n)v_n) = \exp_p(s(v)v). \tag{9.4}$$

Therefore, q is a cut point to p. $\square$

Now Galloway suggested a non-intrinsic way to deal with the nonspacelike cut locus in such a manner as to bring an analogue of (9.4) to bear. The two facts that (i) future inextendible nonspacelike geodesics are not trapped in compact sets in strongly causal space–times, and (ii) a modified s-function defined for all future causal directions to be constructed below is lower semicontinuous for globally hyperbolic space–times suffice to ensure that the basic ideas of the above proof are valid in the space–time setting.

Fix throughout the rest of this section an auxiliary complete Riemannian metric h for the space–time (M, g). Let

$$UM = \{v \in TM : h(v, v) = 1 \text{ and } v \text{ is future directed } \}.$$

Then we define the modified s-function on UM, which will be denoted by s_1, as would be expected.

Definition 9.32. Define the function $s_1 : UM \to \mathbb{R} \cup \{\infty\}$ by

$$s_1(v) = \sup\{t > 0 : d(\pi(v), c_v(t)) = L(c_v \,|\, [0, t])\}$$

where $c_v(t) = \exp(tv)$ as before.

Evidently, for timelike v in UM

$$\|v\| \cdot s_1(v) = s(v/\|v\|), \tag{9.5}$$

and further it is not difficult to see that the semicontinuity proofs given for s and $T_{-1}M$ apply equally well to s_1 and UM with minor modifications. Also, if (M, g) is strongly causal, then $s_1(v) > 0$ for any v in UM.

Proposition 9.33.

(1) *Let $v \in UM$ with $s_1(v) > 0$. Suppose either that $s_1(v) = +\infty$, or $s_1(v)$ is finite and $c_v(t)$ extends to $[0, s_1(v)]$. Then s_1 is upper semicontinuous at v in UM. Especially, if (M, g) is future nonspacelike geodesically complete, then $s_1 : UM \to \mathbb{R} \cup \{\infty\}$ is everywhere upper semicontinuous.*

(2) *If (M, g) is globally hyperbolic, then $s_1 : UM \to \mathbb{R} \cup \{\infty\}$ is lower semicontinuous.*

Before giving the elementary proof of the closure of the nonspacelike cut locus for globally hyperbolic space–times, it is helpful to note the following technical result. It is not required a priori in this result that either of a or b is finite. Similar issues are addressed in a more general context and on a more systematic basis in Lemma 11.20 and the proof of Theorem 11.25.

Lemma 9.34. *Let (M, g) be a globally hyperbolic space–time. Suppose that $c_n : [0, b_n] \to (M, g)$ is a sequence of future nonspacelike maximal geodesic segments with $v_n = c_n'(0)$ in UM and $c_n(0) = p$. Put $q_n = c_n(b_n)$, and suppose that $q_n \to q$ in M, $b_n \to b > 0$ in $\mathbb{R} \cup \{\infty\}$, and $v_n \to v$ in UM. Let $c : [0, a) \to (M, g)$ denote the maximally extended nonspacelike limit geodesic given by $c(t) = \exp_p(tv)$. Then*

(1) *$b < a$, so that b is finite and positive,*

(2) *$c(b) = q$, and*

(3) *$s_1(v) \geq b$.*

Proof. (1) Suppose $a \leq b$. Fix any q' with $q \ll q'$. Then for any s with $0 < s < a$, we have $c_n(s)$ defined for all n sufficiently large, and $c_n(s) \in I^-(q')$ for all such large n. Since $v_n \to v$ and $c_n(s)$ is defined for all sufficiently large n, we have $c(s) = \lim c_n(s)$, whence $c(s) \in J^-(q')$. Thus the future inextendible, nonspacelike geodesic $c \,|\, [0, a)$ is contained in the compact set $J^+(p) \cap J^-(q')$, which is impossible since (M, g) is strongly causal. Thus $b < a$, and $b = \lim b_n$ must be finite as well.

(2) Since $b < a$, we find that $c(b)$ is defined and $q = \lim q_n = \lim c_n(b_n) = c(b)$ by uniform convergence.

(3) Since the geodesic segments $c_n \,|\, [0, b_n]$ are assumed to be maximal and $b = \lim b_n$, the limit segment $c \,|\, [0, b]$ must also be maximal, whence $s_1(v) \geq b$. □

Theorem 9.35. *Let (M, g) be globally hyperbolic. Then the future non-spacelike cut locus and the future null cut locus of any point p in M are closed subsets of M.*

Proof. The proof may be given in a uniform fashion either for $\{q_n\}$ a sequence of null cut points to p or a sequence of nonspacelike cut points to p, so we will not specify. Suppose $q_n \to q$ in M. Choose $v_n \in UM$ with

$$q_n = \exp_p(s_1(v_n)v_n), \tag{9.6}$$

and put $b_n = s_1(v_n)$. Because $\{w \in UM : w \in T_pM\}$ is compact, by taking successive subsequences we may suppose that $v_n \to v$ in UM and that $b_n \to b$ in $\mathbb{R} \cup \{\infty\}$. Put $c(t) = \exp_p(tv)$. By Proposition 9.33–(2), we have $b \geq s_1(v)$, so that $b > 0$. By Lemma 9.34, b is finite, $c(b) = q$, and $s_1(v) \geq b$. Hence $s_1(v) = b$, and thus q is a cut point of the required type. □

G. Galloway has also remarked that with non-intrinsic methods, the hypothesis of "timelike geodesic completeness" may be removed from Proposition 9.30.

Proposition 9.36 (Galloway). *Let (M, g) be globally hyperbolic. Then for each p in M the set $B = J^+(p) - [C^+(p) \cup E^+(p)]$ is an open cell.*

Proof. As before, $B = I^+(p) - C^+(p)$. Fix $c > 0$ sufficiently small that if we set $V = \{v \in TM : h(v, v) = c,\ v \text{ future timelike}\}$, then $U = \exp_p(V)$ is contained in B. Also, fix a diffeomorphism $f : U \to \mathbb{R}^{n-1}$. Define a projection map $\eta : B \to U$ by letting $\eta(m)$ denote the point of intersection of U with the unique future timelike geodesic in M from p to m. Now choose a second auxiliary complete Riemannian metric h_1 just for the open subset B of M, considered as a manifold in its own right. For each q in U, let γ_q denote the g-geodesic $c_{\exp_p^{-1}(q)}$ reparametrized as an h_1-unit speed curve γ_q in B with $\gamma_q(0) = q$ and parametrized with the same orientation as the corresponding g-geodesic. By Lemma 3.65, $\gamma_q : \mathbb{R} \to B$ for each q in U. Now define $t : B \to \mathbb{R}$ by the requirement that

$$\gamma_{\eta(m)}(t(m)) = m.$$

Then we may define a homeomorphism $F : B \to \mathbb{R} \times \mathbb{R}^{n-1}$ by $F(m) = (t(m), f(\eta(m)))$. □

CHAPTER 10

MORSE INDEX THEORY ON LORENTZIAN MANIFOLDS

Given a nonspacelike geodesic $\gamma : [0, a) \to M$ in a causal space–time, we have seen in Chapter 9 that if $\gamma(t_0)$ is the future cut point to $\gamma(0)$ along γ, then for any $t < t_0$ the geodesic segment $\gamma | [0, t]$ is the longest nonspacelike curve from $\gamma(0)$ to $\gamma(t)$ in *all* of M. We could ask a much less stringent question: among all nonspacelike curves σ from $\gamma(0)$ to $\gamma(t_0)$ sufficiently "close" to γ, is $L(\gamma) \geq L(\sigma)$? If so, $\gamma(t_0)$ comes at or before the first future conjugate point of $\gamma(0)$ along γ. The crucial difference here is between "in all of M" for cut points and "close to γ" for conjugate points. The importance of this distinction is illustrated by the fact that while no two-dimensional space–time has any null conjugate points, all null geodesics in the two-dimensional Einstein static universe have future and past null cut points.

Since only the behavior of "nearby" curves is considered in studying conjugate points, it is natural to apply similar techniques from the calculus of variations to geodesics in arbitrary Riemannian manifolds and to nonspacelike geodesics in arbitrary Lorentzian manifolds. To indicate the flavor of the Lorentzian index theory, we sketch the Morse index theory of an arbitrary (not necessarily complete) Riemannian manifold (N, g_0). Let $c : [a, b] \to N$ be a fixed geodesic segment and consider a one-parameter family of curves α_s starting at $c(a)$ and ending at $c(b)$. More precisely, let $\alpha : [a, b] \times (-\epsilon, \epsilon) \to N$ be a continuous, piecewise smooth map with $\alpha(t, 0) = c(t)$ for all $t \in [a, b]$, and $\alpha(a, s) = c(a)$, $\alpha(b, s) = c(b)$ for all $s \in (-\epsilon, \epsilon)$. Thus each neighboring curve $\alpha_s : t \to \alpha_s(t) = \alpha(t, s)$ is piecewise smooth. The *variation vector field* (or *s-parameter derivative*) V of this deformation is given by

$$V(t) = \frac{d}{ds}(s \mapsto \alpha(t, s))\bigg|_{s=0}.$$

Since c is a smooth geodesic, one calculates that

$$\frac{d}{ds}(L_0(\alpha_s))\bigg|_{s=0} = 0,$$

and the second variation works out to be

$$\begin{aligned}\frac{d^2}{ds^2}(L_0(\alpha_s))\bigg|_{s=0} \\ &= \int_{t=a}^{b} [g_0(V', V') - g_0(R(V, c')c', V) - c'(g_0(V, c'))]\big|_t \, dt \\ &\quad + g_0(\nabla_V V, c')\big|_a^b.\end{aligned}$$

This second variation formula naturally suggests defining an *index form*

$$I(X, Y) = \int_{t=a}^{b} [g_0(X', Y') - g_0(R(X, c')c', Y)]\big|_t \, dt$$

on the infinite-dimensional vector space $V_0^\perp(c)$ of piecewise smooth vector fields X along c orthogonal to c with $X(a) = X(b) = 0$. It is then shown that a necessary and sufficient condition for c to be free of conjugate points is that $I(X, X) > 0$ for all nontrivial $X \in V_0^\perp(c)$.

This suggests that for a geodesic segment $c : [a, b] \to N$ with conjugate points to $t = a$, the index $\text{Ind}(c)$ of c with respect to $I : V_0^\perp(c) \times V_0^\perp(c) \to \mathbb{R}$ should be defined as the supremum of dimensions of all vector subspaces of $V_0^\perp(c)$ on which I is negative definite. Even though $V_0^\perp(c)$ is an infinite-dimensional vector space, the Morse Index Theorem for arbitrary Riemannian manifolds asserts that:

(1) $\text{Ind}(c)$ is finite; and

(2) $\text{Ind}(c)$ equals the geodesic index of c, i.e., the number of conjugate points along c counted with multiplicities.

More precisely, if we let $J_t(c)$ denote the vector space of smooth vector fields Y along c satisfying the Jacobi differential equation $Y'' + R(Y, c')c' = 0$ with boundary conditions $Y(a) = Y(t) = 0$, then (2) is equivalent to the formula

(3) $\displaystyle \text{Ind}(c) = \sum_{t \in (a,b)} \dim J_t(c).$

Also, c has only finitely many points in $(a, b]$ conjugate to $c(a)$.

With the Morse Index Theorem in hand, the homotopy type of the loop space for *complete* Riemannian manifolds may now be calculated geometrically [cf. Milnor (1963, p. 95)]. One obtains the result that if (N, g_0) is a complete Riemannian manifold and $p, q \in N$ are any pair of points which are not conjugate along any geodesic, then the loop space $\Omega_{(p,q)}$ of all continuous paths from p to q equipped with the compact-open topology has the homotopy type of a countable CW-complex which contains a cell of dimension λ for each geodesic from p to q of index λ.

The purpose of Sections 10.1 and 10.3 is to prove the analogues of (1) through (3) for nonspacelike geodesics in arbitrary space–times. Let $c : [a, b] \to M$ [respectively, $\beta : [a, b] \to M$] denote an arbitrary timelike [respectively, null] geodesic segment in (M, g). Let $V_0^\perp(c)$ [respectively, $V_0^\perp(\beta)$] denote the infinite-dimensional vector space of piecewise smooth vector fields Y along c [respectively, β] perpendicular to c [respectively, β] with $Y(a) = Y(b) = 0$. The timelike index form $I : V_0^\perp(c) \times V_0^\perp(c) \to \mathbb{R}$ may be defined by

$$I(X, Y) = -\int_a^b [g(X', Y') - g(R(X, c')c', Y)]dt$$

in analogy with the Riemannian index form. It may also be shown that $c : [a, b] \to M$ has no conjugate points in (a, b) if and only if $I : V_0^\perp(c) \times V_0^\perp(c) \to \mathbb{R}$ is negative definite.

Similarly, an index form $I : V_0^\perp(\beta) \times V_0^\perp(\beta) \to \mathbb{R}$ may be defined by $I(X, Y) = -\int_a^b [g(X', Y') - g(R(X, \beta')\beta', Y)]\, dt$. But since β is a null geodesic, $g(\beta', \beta') = 0$. Consequently, vector fields of the form $V(t) = f(t)\beta'(t)$ with $f : [a, b] \to \mathbb{R}$ piecewise smooth and $f(a) = f(b) = 0$ are always in the null space of $I : V_0^\perp(\beta) \times V_0^\perp(\beta) \to \mathbb{R}$ yet never give rise to null conjugate points. One way around this difficulty is to consider the quotient bundle $\mathfrak{X}_0(\beta)$ of $V_0^\perp(\beta)$ formed by identifying Y_1 and Y_2 in $V_0^\perp(\beta)$ if $Y_1 - Y_2 = f\beta'$ for some piecewise smooth function $f : [a, b] \to \mathbb{R}$. The index form $I : V_0^\perp(\beta) \times V_0^\perp(\beta) \to \mathbb{R}$ may also be projected to a quotient index form $\overline{I} : \mathfrak{X}_0(\beta) \times \mathfrak{X}_0(\beta) \to \mathbb{R}$. It may then be shown that the null geodesic segment $\beta : [a, b] \to M$ has no conjugate points in $[a, b]$ if and only if

$\overline{I} : \mathfrak{X}_0(\beta) \times \mathfrak{X}_0(\beta) \to \mathbb{R}$ is negative definite [cf. Hawking and Ellis (1973, Section 4.5), Bölts (1977, Chapters 2 and 4)]. In the case of a congruence of null geodesics, the procedure of forming the quotient bundle has also been utilized in general relativity since the 1960's, and the term "screen space" has been employed [cf. Robinson and Trautman (1983) for a recent reference].

Let $J_t(c)$ [respectively, $J_t(\beta)$] denote the space of Jacobi fields along c [respectively, β] with $Y(a) = Y(b) = 0$. Define the *index* $\operatorname{Ind}(c)$ of c [respectively, $\operatorname{Ind}(\beta)$ of β] to be the supremum of dimensions of all vector subspaces of $V_0^{\perp}(c)$ [respectively, $\mathfrak{X}_0(\beta)$] on which I [respectively, $\overline{I}$] is positive definite. We establish the Morse Index Theorem

$$\operatorname{Ind}(c) = \sum_{t \in (a,b)} \dim J_t(c)$$

and

$$\operatorname{Ind}(\beta) = \sum_{t \in (a,b)} \dim J_t(\beta)$$

for the timelike geodesic $c : [a, b] \to M$ and the null geodesic $\beta : [a, b] \to M$ in Sections 10.1 and 10.3, respectively. The reason for considering the timelike and null cases separately is that the quotient index form $\overline{I} : \mathfrak{X}_0(\beta) \times \mathfrak{X}_0(\beta) \to \mathbb{R}$ must be used to obtain the Null, but not the Timelike, Morse Index Theorem.

In Section 10.2 we study the theory of timelike loop spaces for globally hyperbolic space–times, summarizing some results of Uhlenbeck (1975). Since completeness is necessary to develop the Riemannian loop space theory, it is not surprising that global hyperbolicity is needed for the Lorentzian theory.

Yet a significant difference occurs between the Lorentzian and Riemannian loop spaces. Let (N, g_0) be a complete Riemannian manifold with positive Ricci curvature bounded away from zero. Thus N is compact by Myers' Theorem. It is shown in Milnor (1963, Theorem 19.6) that if p and q in such a Riemannian manifold are nonconjugate along any geodesic, then the loop space $\Omega_{(p,q)}$ has the homotopy type of a CW-complex having only finitely many cells in each dimension. But the loop space may still be an *infinite* CW-complex as is seen by the example of S^n with the usual complete metric of constant sectional curvature. On the other hand, if (M, g) is an *arbitrary* globally hyperbolic (hence noncompact) space–time, and $p, q \in M$ are any two points with $p \ll q$ such that p and q are not conjugate along any nonspacelike geodesic, then the timelike loop space $C_{(p,q)}$ has only *finitely* many cells.

This striking difference between the Lorentzian and Riemannian loop spaces is a result of the following basic difference between Lorentzian and Riemannian manifolds. If γ is any nonspacelike curve from p to q in an arbitrary space–time and $d(p,q)$ is finite, then γ has bounded Lorentzian arc length since $L(\gamma) \leq d(p,q)$. In contrast, any curve σ in the Riemannian path space satisfies $L_0(\sigma) \geq d_0(p,q)$, hence has arc length bounded from below but not above.

10.1 The Timelike Morse Index Theory

In this section we give a detailed proof of the Morse Index Theorem for timelike geodesics in an arbitrary space–time. Several different approaches to Morse index theory for timelike geodesics under various causality conditions have been published by Uhlenbeck (1975), Woodhouse (1976), Everson and Talbot (1976), and Beem and Ehrlich (1979c). Here we give a treatment which parallels the proof of the Morse Index Theorem for an arbitrary Riemannian manifold in Gromoll, Klingenberg, and Meyer (1975, Sections 4.5 and 4.6). A similar treatment of the results in this section through Theorem 10.22 has been given in Bölts (1977).

Let (M,g) be an *arbitrary* space–time of dimension $n \geq 2$ throughout this section. We will denote by $\langle \;,\; \rangle$ the Lorentzian metric g in this section. Also, all timelike geodesics $c : [a,b] \to M$ will be assumed to have unit speed, i.e., $\langle c'(t), c'(t) \rangle = -1$ for all $t \in [a,b]$.

Definition 10.1. (*Piecewise Smooth Vector Field*) A *piecewise smooth vector field Y along* (the timelike geodesic) $c : [a,b] \to M$ is a continuous map $Y : [a,b] \to TM$, with $Y(t) \in T_{c(t)}M$ for all $t \in [a,b]$, such that there exists some finite partition $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ of $[a,b]$ so that $Y|_{[t_i,t_{i+1}]} : [t_i, t_{i+1}] \to TM$ is smooth for each i with $0 \leq i \leq k-1$. Let $V^\perp(c)$ denote the $\mathbb{R}$-vector space of all piecewise smooth vector fields Y along c with $\langle Y(t), c'(t) \rangle = 0$ for all $0 \leq t \leq b$. Also let $V_0^\perp(c) = \{Y \in V^\perp(c) : Y(a) = Y(b) = 0\}$ and $N(c(t)) = \{v \in T_{c(t)}M : \langle v, c'(t) \rangle = 0\}$.

Given $Y \in V^{\perp}(c)$, there is a finite set of points $I_0 \subseteq (a,b)$ such that Y is differentiable on $[a,b] - I_0$. Define $Y' : [a,b] \to TM$ by $Y'(t) = \nabla_{\gamma'} Y(t)$ for $t \in [a,b] - I_0$ and extend to points $t_i \in I_0$ by setting

$$Y'(t_i) = \lim_{t \to t_i^-} Y'(t).$$

Thus Y' is extended to $[a,b]$ by left continuity but is not necessarily continuous at points of I_0.

Remark 10.2. Since $c'(t)$ is timelike, $N(c(t))$ is a vector space of dimension $n-1$ consisting of spacelike tangent vectors, and thus $\{v \in N(c(t)) : \langle v, v \rangle \leq 1\}$ is compact.

Definition 10.3. (*Conjugate Point*) Let $c : [a,b] \to M$ be a timelike geodesic. Then $t_1, t_2 \in [a,b]$ with $t_1 \neq t_2$ are *conjugate with respect to* c if there is a nontrivial smooth vector field J along c with $J(t_1) = J(t_2) = 0$ satisfying the Jacobi equation $J'' + R(J, c')c' = 0$. Here J' denotes the covariant derivative operator on vector fields along c induced by the Levi–Civita connection of $\langle \ , \ \rangle$ on M. Also $t_1 \in (a,b)$ is said to be a *conjugate point* of the geodesic segment $c : [a,b] \to M$ if a and t_1 are conjugate along c. The geodesic segment c is said to have *no conjugate points* if no $t_1 \in (a,b]$ is conjugate to $t = a$ along c. A smooth vector field J along c satisfying the differential equation $J'' + R(J,c')c' = 0$ is called a *Jacobi field along* c.

We now define the Lorentzian index form $I : V^{\perp}(c) \times V^{\perp}(c) \to \mathbb{R}$.

Definition 10.4. (*Lorentzian Index Form*) The *index form* $I : V^{\perp}(c) \times V^{\perp}(c) \to \mathbb{R}$ is the symmetric bilinear form given by

$$I(X,Y) = -\int_a^b [\langle X', Y' \rangle - \langle R(X,c')c', Y \rangle] \, dt \tag{10.1}$$

for any $X, Y \in V^{\perp}(c)$.

If $X \in V^{\perp}(c)$ is smooth, we also have

$$I(X,Y) = -\left. \langle X', Y \rangle \right|_a^b + \int_a^b \langle X'' + R(X,c')c', Y \rangle \, dt. \tag{10.2}$$

Thus if $Y \in V_0^\perp(c)$ and X is smooth, one obtains the formula

$$I(X,Y) = \int_a^b \langle X'' + R(X,c')c', Y\rangle \, dt \tag{10.3}$$

linking the index form to Jacobi fields.

Remark 10.5. Given a piecewise smooth vector field $X \in V^\perp(c)$, we may choose a partition $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ such that $X \mid [t_i, t_{i+1}]$ is smooth for each $i = 0, 1, 2, \ldots, k-1$. Let

$$\begin{aligned} \Delta_{t_0}(X') &= X'(a^+), \\ \Delta_{t_k}(X') &= -X'(b^-), \quad \text{and} \\ \Delta_{t_i}(X') &= \lim_{t \to t_i^+} X'(t) - \lim_{t \to t_i^-} X'(t) \end{aligned}$$

for $i = 1, 2, \ldots, k-1$. It may then be seen by applying (10.2) on each subinterval (t_i, t_{i+1}) that

$$I(X,Y) = \sum_{i=0}^{k} \langle \Delta_{t_i}(X'), Y\rangle + \int_a^b \langle X'' + R(X,c')c', Y\rangle dt. \tag{10.4}$$

This is the form of the second variation formula given in Hawking and Ellis (1973, p. 108) and Bölts (1977, pp. 86–87) [cf. Cheeger and Ebin (1975, p. 21)].

In order to give geometric applications using the index form, it is useful to make the following standard definition.

Definition 10.6. (*Variation*) Let $c : [a,b] \to M$ be a smooth curve. A *variation* (or homotopy) of c is a smooth mapping $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$, for some $\epsilon > 0$, with $\alpha(t,0) = c(t)$ for all $t \in [a,b]$. The variation α is said to be a *proper variation* if $\alpha(a,s) = c(a)$ and $\alpha(b,s) = c(b)$ for all $s \in (-\epsilon, \epsilon)$. A continuous mapping $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ is said to be a *piecewise smooth variation* of c if there exists a finite partition $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ of $[a,b]$ such that $\alpha \mid [t_i, t_{i+1}] \times (-\epsilon, \epsilon)$ is a smooth variation of $c \mid [t_i, t_{i+1}]$ for each $i = 0, 1, 2, \ldots, k-1$.

If we set $\alpha_s(t) = \alpha(t,s)$ in Definition 10.6, then for a smooth variation α, each curve $\alpha_s : [a,b] \to M$ is a smooth curve and thus the mapping $s \to \alpha_s$

is a deformation of the curve c through the "neighboring curves" α_s. If α is a piecewise smooth variation, the neighboring curves α_s will be piecewise smooth. If α is a proper variation, all of the neighboring curves α_s begin at $c(a)$ and end at $c(b)$. It is customary in defining variations of timelike curves to restrict consideration to variations having the additional property that all neighboring curves $\alpha_s : [a,b] \to M$ are timelike. However, if we use Definition 10.6, this restriction is unnecessary by the following lemma.

Lemma 10.7. *Let $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ be a piecewise smooth variation of the timelike geodesic segment $c : [a,b] \to M$. Then there exists a constant $\delta > 0$ depending on α such that the neighboring curves α_s are timelike for all s with $|s| \leq \delta$.*

Proof. We first suppose that α is a smooth variation. Choose any ϵ_1 with $0 < \epsilon_1 < \epsilon$. Then α is differentiable on the compact set $[a,b] \times [-\epsilon_1, \epsilon_1]$ by Definition 10.6. Hence by definition of differentiability, α extends to a smooth mapping of a larger open set containing $[a,b] \times [-\epsilon_1, \epsilon_1]$. Since c is a timelike geodesic, the vectors $c'(a^+)$ and $c'(b^-)$ are timelike. It follows from this and the extension of α to an open set containing $[a,b] \times [-\epsilon_1, \epsilon_1]$ that there exists a constant $\delta_1 > 0$ such that the tangent vectors ${\alpha_s}'(a^+)$ and ${\alpha_s}'(b^-)$ to the neighboring curves α_s are timelike for all s with $|s| < \delta_1$.

Suppose now that no $\delta > 0$ can be found such that all the curves α_s are timelike for $|s| < \delta$. Then we could find a sequence $s_n \to 0$ such that the curves α_{s_n} failed to be timelike. Thus there would be $t_n \in [a,b]$ so that $g\left(\alpha'_{s_n}(t_n), \alpha'_{s_n}(t_n)\right) \geq 0$ for each n. Since $[a,b] \times [-\epsilon_1, \epsilon_1]$ is compact, the sequence $\{(t_n, s_n)\}$ has a point of accumulation (t, s). Since $s_n \to 0$, this point must be of the form $(t, 0)$, and also the existence of δ_1 above shows that $t \neq a, b$. But then as $g\left(\alpha'_{s_n}(t_n), \alpha'_{s_n}(t_n)\right) \geq 0$ for each n, it follows that $g(c'(t), c'(t)) \geq 0$, in contradiction to the fact that c was a timelike geodesic segment. Thus we have seen that if $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ is a smooth variation of the timelike geodesic $c : [a,b] \to M$, then there is a constant $\delta > 0$

such that $\alpha_s'(a^+)$ and $\alpha_s'(b^-)$ are timelike vectors for all $|s| \leq \delta$ and the curves α_s are timelike for all $|s| \leq \delta$.

Now let $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ be a piecewise smooth variation of the timelike geodesic $c : [a,b] \to M$. There is by Definition 10.6 a finite partition $a = t_0 < t_1 < \cdots < t_k = b$ such that $\alpha \,|\, [t_i, t_{i+1}] \times (-\epsilon, \epsilon) \to M$ is a smooth variation of $c \,|\, [t_i, t_{i+1}]$. By the above paragraph, we may find a constant $\delta_i > 0$ such that for all s with $|s| \leq \delta_i$, the tangent vectors $\alpha_s'(t_i^+)$ and $\alpha_s'(t_{i+1}^-)$ are timelike and $\alpha_s \,|\, [t_i, t_{i+1}]$ is a timelike curve. Taking $\delta = \min\{\delta_0, \delta_1, \ldots, \delta_{k-1}\}$ then yields the required δ. □

Remark 10.8. There is no result corresponding to Lemma 10.7 for variations of null geodesics (cf. Definition 10.58 ff.).

The index form may now be related to variations of timelike geodesic segments $c : [a,b] \to M$ as follows. Given $Y \in V_0^\perp(c)$, define the *canonical proper variation* $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ of c by setting

$$\alpha(t,s) = \exp_{c(t)}(sY(t)). \tag{10.5}$$

It should first be noted that since $c([a,b])$ is a compact subset of M, and the differential of the exponential map, $\exp_{p_*}$, is nonsingular at the origin of T_pM for all $p \in M$, it is possible given $c([a,b])$ to find an $\epsilon > 0$ such that $\exp_{c(t)}(sY(t))$ is defined for all s with $|s| \leq \epsilon$ and for each $t \in [a,b]$. Secondly, from Definition 10.1 it follows that $\alpha(t,s)$ defined as in (10.5) is a piecewise smooth variation of c. Hence given $Y \in V_0^\perp(c)$, we know from Lemma 10.7 that there exists some constant $\delta > 0$ such that all the neighboring curves $\alpha_s : t \to \alpha(t,s)$ are timelike for all s with $-\delta < s < \delta$.

Given an arbitrary smooth variation $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ of $c : [a,b] \to M$, the *variation vector field* V of α is defined to be the vector field $V(t)$ along c given by the formula

$$V(t) = \frac{d}{ds}(\alpha(t,s))\bigg|_{s=0}. \tag{10.6}$$

More precisely, letting $\partial/\partial s$ be the coordinate vector field on $[a,b] \times (-\epsilon, \epsilon)$ corresponding to the s parameter, the variation vector field is given by

$$V(t) = \alpha_* \left.\frac{\partial}{\partial s}\right|_{(t,0)}. \tag{10.7}$$

For a piecewise smooth variation $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$, one obtains a continuous piecewise smooth variation vector field as follows. Let $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ be a partition of $[a,b]$ such that $\alpha | [t_i, t_{i+1}] \times (-\epsilon, \epsilon)$ is smooth for $i = 0, 1, \ldots, k-1$. Given $t \in [a,b]$, choose an index i such that $t_i \leq t < t_{i+1}$, and set

$$V(t) = [\alpha | [t_i, t_{i+1}] \times (-\epsilon, \epsilon)]_* \left. \frac{\partial}{\partial s} \right|_{(t,0)} .$$

The canonical variation (10.5) has the property that each curve $s \to \alpha(t,s)$ is just the geodesic $s \to \exp_{c(t)}(sY(t))$ which has initial direction $Y(t)$ at $s = 0$. Hence the variation vector field of the canonical variation (10.5) is just the given vector field $Y \in V_0^{\perp}(c)$.

If we put $L(s) = L(\alpha_s) = L(t \to \alpha(t,s))$, then $L'(0) = \left. \frac{dL(s)}{ds} \right|_{s=0} = 0$ since c is a smooth timelike geodesic, and

$$L''(0) = \left. \frac{d^2}{ds^2} L(s) \right|_{s=0} = I(Y,Y). \tag{10.8}$$

Thus if $I(Y,Y) > 0$ for some $Y \in V_0^{\perp}(c)$, then the canonical proper variation $\alpha(t,s)$ defined by (10.5) using Y will produce timelike neighboring curves α_s joining $c(a)$ to $c(b)$ with $L(\alpha_s) > L(c)$ for s sufficiently small. Thus if the timelike geodesic $c : [a,b] \to M$ is maximal [i.e., $L(c) = d(p,q)$], then $I : V_0^{\perp}(c) \times V_0^{\perp}(c) \to \mathbb{R}$ must be negative semidefinite. Before proving the Morse Index Theorem for timelike geodesics, we must establish the following more precise relationship between conjugate points, Jacobi fields, and the index form. First, the null space of the index form on $V_0^{\perp}(c)$ consists of the *smooth* Jacobi fields in $V_0^{\perp}(c)$, and second, c has no conjugate points on $[a,b]$ if and only if the index form is negative definite on $V_0^{\perp}(c)$.

We first derive an elementary but important consequence of the Jacobi differential equation.

Lemma 10.9. *Let $c : [a,b] \to M$ be a timelike geodesic segment and let Y be any Jacobi field along c. Then $\langle Y(t), c'(t) \rangle$ is an affine function of t, i.e., $\langle Y(t), c'(t) \rangle = \alpha t + \beta$ for some constants $\alpha, \beta \in \mathbb{R}$.*

Proof. First $\langle Y, c' \rangle' = \langle Y', c' \rangle + \langle Y, c'' \rangle = \langle Y', c' \rangle$, as $c'' = \nabla_{c'} c' = 0$ along the geodesic c. Differentiating again, we obtain $\langle Y, c' \rangle'' = \langle Y'', c' \rangle +$

$\langle Y', c''\rangle = \langle Y'', c'\rangle = -\langle R(Y, c')c', c'\rangle = 0$ by the skew symmetry of the Riemann–Christoffel tensor. Thus $\langle Y, c'\rangle$ is an affine function. $\square$

Corollary 10.10. *If Y is any Jacobi field along the timelike geodesic $c : [a, b] \to M$ and $Y(t_1) = Y(t_2) = 0$ for distinct $t_1, t_2 \in [a, b]$, then $Y \in V^\perp(c)$.*

Corollary 10.11. *If $Y \in V_0^\perp(c)$ is a Jacobi field, then $Y' \in V^\perp(c)$.*

Using the canonical variation, we are now ready to derive the following geometric consequence of the existence of a conjugate point $t_0 \in (a, b)$ to $t = a$ along c.

Proposition 10.12. *Suppose that the timelike geodesic $c : [a, b] \to M$ contains a conjugate point $t_0 \in (a, b)$ to $t = a$ along c. Then there exists a piecewise smooth proper variation α_s of c such that $L(\alpha_s) > L(c)$ for all $s \neq 0$. Thus $c : [a, b] \to M$ is not maximal.*

Proof. In view of (10.5), (10.8), and Lemma 10.7, it is enough to construct a piecewise smooth vector field $Y \in V_0^\perp(c)$ with $I(Y, Y) > 0$ and let α be the canonical variation associated with Y. To this end, let Y_1 be a nontrivial Jacobi field along c with $Y_1(a) = Y_1(t_0) = 0$. By Corollary 10.10, $Y_1 \in V^\perp(c)$. Hence as $Y_1(a) = Y_1(t_0) = 0$, we have $Y_1' \in V^\perp(c)$ by Corollary 10.11. Since $Y_1(t_0) = 0$ and Y_1 is a nontrivial Jacobi field, it follows that $Y_1'(t_0)$ is a (nonzero) spacelike tangent vector.

Let $I(\ ,\ )_a^s$ denote the restriction of the index form to $c \mid [a, s]$, that is,

$$I(V, W)_a^s = -\int_a^s [\langle V', W'\rangle - \langle R(V, c')c', W\rangle]\, dt.$$

Then since Y_1 is a Jacobi field, we have from (10.2) of Definition 10.4 that

$$I(Y_1, Z)_a^s = -\langle Y_1', Z\rangle|_a^s \tag{10.9}$$

for any $Z \in V^\perp(c)$.

We are now ready to construct a piecewise smooth vector field $Y \in V_0^\perp(c)$ with $I(Y, Y) > 0$. Let $\psi : [a, b] \to \mathbb{R}$ be a smooth function with $\psi(a) = \psi(b) = 0$ and $\psi(t_0) = 1$. Also let Z_1 be the unique smooth parallel vector field along

c with $Z_1(t_0) = -Y_1{}'(t_0)$. Then $Z = \psi Z_1 \in V_0^\perp(c)$. Define a one-parameter family $Y_\epsilon \in V_0^\perp(c)$ by

$$Y_\epsilon(t) = \begin{cases} Y_1(t) + \epsilon Z(t) & \text{if } a \le t \le t_0, \\ \epsilon Z(t) & \text{if } t_0 \le t \le b. \end{cases}$$

Then using (10.9) we obtain

$$\begin{aligned} I(Y_\epsilon, Y_\epsilon) &= I(Y_\epsilon, Y_\epsilon)_a^{t_0} + I(Y_\epsilon, Y_\epsilon)_{t_0}^b \\ &= I(Y_1 + \epsilon Z, Y_1 + \epsilon Z)_a^{t_0} + I(\epsilon Z, \epsilon Z)_{t_0}^b \\ &= I(Y_1, Y_1)_a^{t_0} + 2\epsilon I(Y_1, Z)_a^{t_0} + \epsilon^2 I(Z, Z)_a^{t_0} + \epsilon^2 I(Z, Z)_{t_0}^b \\ &= -\left\langle Y_1{}', Y_1 \right\rangle\Big|_a^{t_0} - 2\epsilon \left\langle Y_1{}', Z \right\rangle\Big|_a^{t_0} + \epsilon^2 I(Z, Z). \end{aligned}$$

Since $Y_1(a) = Y_1(t_0) = 0$, this simplifies to

$$\begin{aligned} I(Y_\epsilon, Y_\epsilon) &= -2\epsilon \langle Y_1{}'(t_0), Z(t_0) \rangle + \epsilon^2 I(Z, Z) \\ &= 2\epsilon \parallel Y_1{}'(t_0) \parallel^2 + \epsilon^2 I(Z, Z). \end{aligned}$$

As $Y_1{}'(t_0)$ is a (nonzero) spacelike tangent vector and $I(Z, Z)$ is finite, it follows that there is some $\epsilon > 0$ such that $I(Y_\epsilon, Y_\epsilon) > 0$. Put the required vector field $Y = Y_\epsilon$ for this value of ϵ. $\square$

We now turn to an important characterization of Jacobi fields in terms of the index form. The same characterization holds for Riemannian spaces with an identical proof. It is important to note that the index form characterizes smooth Jacobi fields among all *piecewise smooth* vector fields in $V_0^\perp(c)$ and not just among all smooth vector fields in $V_0^\perp(c)$.

Proposition 10.13. *Let $c : [a, b] \to (M, g)$ be a timelike geodesic segment. Then for $Y \in V_0^\perp(c)$, the following are equivalent:*

(1) *Y is a (smooth) Jacobi field along c.*

(2) *$I(Y, Z) = 0$ for all $Z \in V_0^\perp(c)$.*

Proof. First (1) $\Rightarrow$ (2) is immediate, since for smooth vector fields Y and arbitrary Z the index form may be written as

$$I(Y, Z) = -\left\langle Y', Z \right\rangle\Big|_a^b + \int_a^b \langle Y'' + R(Y, c')c', Z \rangle dt.$$

If Y is a Jacobi field, $I(Y,Z) = -\langle Y', Z\rangle|_a^b$. Thus $I(Y,Z)$ vanishes for all $Z \in V_0^\perp(c)$.

To show (2) $\Rightarrow$ (1), we first note that since c is a timelike geodesic segment and $Y \in V_0^\perp(c)$, we have $\langle Y', c'\rangle = 0$ and $\langle Y'' + R(Y,c')c', c'\rangle = 0$ at all points where Y is differentiable. Taking left-hand limits, we have $\langle Y'(t_i), c'(t_i)\rangle = 0$ also at the finitely many points of discontinuity $a = t_0 < t_1 < \cdots < t_k = b$ of Y. By continuity, the right-hand limit $\lim_{t\to t_i^+}\langle Y'(t), c'(t)\rangle = 0$ as well. Hence the vectors $\Delta_{t_i}(Y')$ defined in Remark 10.5 also satisfy $\langle \Delta_{t_i}(Y'), c'(t_i)\rangle = 0$. By (10.4) of Remark 10.5, the index form may be calculated as

$$I(Y,Z) = \sum_{i=0}^{k} \langle \Delta_{t_i}(Y'), Z(t_i)\rangle + \int_a^b \langle Y'' + R(Y,c')c', Z\rangle\, dt. \tag{10.10}$$

Let $\phi : [a,b] \to [0,1]$ be a smooth function with $\phi(t_0) = \phi(t_1) = \cdots = \phi(t_k) = 0$ and $\phi(t) > 0$ elsewhere. Then the vector field $Z_1 = \phi(Y'' + R(Y,c')c')$ is in $V_0^\perp(c)$ and $Z_1(t_i) = 0$ for all i. Since it is assumed that $I(Y,Z) = 0$ for all $Z \in V_0^\perp(c)$, we obtain from (10.10) that

$$0 = I(Y, Z_1) = \int_a^b \phi(t)\, \|Y'' + R(Y,c')c'\|^2\,\big|_t\, dt.$$

As Z_1 is a spacelike vector field, smooth except at the t_i's, and $\phi(t) > 0$ if $t \notin \{t_i\}$, we obtain $Y''(t) + R(Y(t), c'(t))c'(t) = 0$ if $t \notin \{t_i\}$. Thus Y is a piecewise Jacobi field and formula (10.10) reduces to

$$I(Y,Z) = \sum_{i=0}^{k} \langle \Delta_{t_i}(Y'), Z(t_i)\rangle. \tag{10.11}$$

Recalling from above that $\langle \Delta_{t_i}(Y'), c'(t_i)\rangle = 0$ for each i, a vector field $Z_2 \in V_0^\perp(c)$ may be constructed with $Z_2(t_i) = \Delta_{t_i}(Y')$ for $i = 1, 2, \ldots, k-1$. Then we have

$$0 = I(Y, Z_2) = \sum_{i=1}^{k-1} \|\Delta_{t_i}(Y')\|^2.$$

Since all of the tangent vectors in this sum are spacelike, it follows that $\Delta_{t_i}(Y') = 0$ for $i = 1, 2, \ldots, k-1$. This then implies that Y' has no breaks at the t_i's. Since for any $t \in [a,b]$ there is a unique Jacobi field along c

with $Y(t) = v$ and $Y'(t) = w$, it follows that the Jacobi fields $Y \mid [t_i, t_{i+1}]$ fit together to form a smooth Jacobi field. □

In view of Propositions 10.12 and 10.13, it should come as no surprise that the negative definiteness of the Lorentzian index form should be related to the absence of conjugate points just as the positive definiteness of the Riemannian index form is guaranteed by the nonexistence of conjugate points [Gromoll, Klingenberg, and Meyer (1975, p. 145)]. The negative semidefiniteness of the index form in the absence of conjugate points has been given in Hawking and Ellis (1973, Lemma 4.5.8). It has been noted in Bölts (1977, Satz 4.4.5) and Beem and Ehrlich (1979c, p. 376) that the negative definiteness of the index form in the absence of conjugate points follows "algebraically" from the semidefiniteness just as in the proof of positive definiteness for the Riemannian index form. In order to give a proof of the negative semidefiniteness of the Lorentzian index form in the absence of conjugate points, we need to obtain the Lorentzian analogues of several important results in Riemannian geometry [cf. Gromoll, Klingenberg, and Meyer (1975, pp. 132, 136–137, 140)].

For the purpose of constructing Jacobi fields, it is useful to introduce some notation for the identification of the tangent space $T_v(T_pM)$ with T_pM itself by "parallel translation in T_pM."

Definition 10.14. (*Tangent Space* $T_v(T_pM)$) Given any $p \in M$ and $v \in T_pM$, the *tangent space* $T_v(T_pM)$ *to the tangent space* T_pM *at* v is given by

$$T_v(T_pM) = \{\phi_w : \mathbb{R} \to T_pM\}$$

where

$$\phi_w(t) = v + tw.$$

Then $T_v(T_pM)$ may intuitively be identified with T_pM by identifying the image of ϕ_w in T_pM with the vector w. More formally, let $\mathbb{R}$ be given the usual manifold coordinate chart $x(r) = r$ for all $r \in \mathbb{R}$, i.e., $x = \text{id}$. Then $\partial/\partial x$ is a vector field on $\mathbb{R}$. Since $\phi_w : \mathbb{R} \to T_pM$ and $\phi_w(0) = v$, we have $\phi_{w_*} : T_0\mathbb{R} \to T_v(T_pM)$. We may then make the following definition.

Definition 10.15. (*Canonical Isomorphism*) The *canonical isomorphism* $\tau_v : T_pM \to T_v(T_pM)$ is given by

$$\tau_v w = \phi_{w*} \left.\frac{\partial}{\partial x}\right|_0 = \phi_w{}'(0)$$

where, as in Definition 10.14,

$$\phi_w(t) = v + tw.$$

In particular, let $v = 0_p$, the zero vector in the tangent space T_pM. Then $\phi_w : \mathbb{R} \to T_pM$ is the curve $\phi_w(t) = tw$ in $T_{0_p}(T_pM)$, and we find that $\tau_{0_p}(w) = \phi_w$. Thus $T_{0_p}(T_pM)$ is often canonically identified with T_pM itself by identifying the vector $w \in T_pM$ and the map ϕ_w. If $p \in M$ and $v \in T_pM$, then since $\exp_p : T_pM \to M$, the definition of the differential gives

$$\exp_{p_*} : T_v(T_pM) \to T_{\exp_p(v)}M.$$

In particular, for $v = 0_p$ we have

$$\exp_{p_*} : T_{0_p}(T_pM) \to T_pM$$

since $\exp_p(0_p) = p$. If $b = \tau_{0_p}(v) \in T_{0_p}(T_pM)$ and we define $\phi : \mathbb{R} \to T_pM$ by $\phi(t) = tv$, then $\exp_{p_*}(b) = \exp_{p_*}(\phi_* \,\partial/\partial x|_0)$, where $\partial/\partial x$ is the usual basis vector field for $T\mathbb{R}$ defined above. Thus we obtain

$$\begin{aligned}\exp_{p_*}(b) &= (\exp_{p_*} \circ \phi_*)\left(\left.\frac{\partial}{\partial x}\right|_0\right) = (\exp_p \circ \phi)_*\left(\left.\frac{\partial}{\partial x}\right|_0\right) \\ &= \left.\frac{d}{dt}(t \to \exp_p(tv))\right|_{t=0} = v.\end{aligned}$$

Thus $\exp_{p_{*0_p}} \circ \tau_v = \mathrm{Id}_{T_pM}$. This fact is commonly stated as "the differential of the exponential map at the origin of T_pM is the identity."

The following proposition shows how the differential of the exponential map may be used to construct Jacobi fields.

Proposition 10.16. *Let $c : [0, b] \to M$ be a geodesic with $c(0) = p$. Let $w \in T_pM$ be arbitrary. Then the unique Jacobi field J along c with $J(0) = 0$ and $J'(0) = w$ is given by*

$$J(t) = \exp_{p_*}(t\,\tau_{tc'(0)}w).$$

Proof. Set $v = c'(0)$. We may find an $\epsilon > 0$ so that the smooth variation $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ of $c : [a, b] \to M$ given by $\alpha(t, s) = \exp_p(t(v + sw))$ is defined. Since this is a variation of c whose s-parameter curves $t \to \alpha(t, s)$ are geodesics, it follows that the variation vector field of this deformation is a Jacobi field. Since $\alpha_* \,\partial/\partial s|_{(t,s)} = \exp_{p_*}(\tau_{t(v+sw)}(tw))$, the variation vector field is just $J(t) = \exp_{p_*}(\tau_{tv}tw) = \exp_{p_*}(t\,\tau_{tv}w)$. Since $\alpha(0, s) = c(0)$ for all s, we have $J(0) = 0$, and a calculation also gives $J'(0) = w$ [cf. Gromoll, Klingenberg, and Meyer (1975, p. 132)]. □

As in the Riemannian theory, it is now possible to prove the Gauss Lemma using Proposition 10.16. We first need to put a natural inner product $\langle\langle\ ,\ \rangle\rangle$ on $T_v(T_pM)$ using the given Lorentzian metric $\langle\ ,\ \rangle$ on T_pM and the canonical isomorphism.

Definition 10.17. (*Inner Product on $T_v(T_pM)$*) The *inner product*

$$\langle\langle\ ,\ \rangle\rangle : T_v(T_pM) \times T_v(T_pM) \to \mathbb{R}$$

associated with the Lorentzian metric $\langle\ ,\ \rangle$ for M is given by

$$\langle\langle a, b\rangle\rangle = \langle \tau_v^{-1}(a), \tau_v^{-1}(b)\rangle$$

for any $a, b \in T_v(T_pM)$.

Theorem 10.18 (Gauss Lemma). *Let $v \in T_pM$ be a tangent vector in the domain of definition of the exponential mapping and let $a = \tau_v(v) \in T_v(T_pM)$. Then for any $b \in T_v(T_pM)$, we have*

$$\langle\langle a, b\rangle\rangle = \langle \exp_{p_*} a, \exp_{p_*} b\rangle. \tag{10.12}$$

Thus the exponential map is a "radial isometry."

Proof. If $\phi(t) = tv$, then $a = \phi'(1)$. If c is the geodesic $c(t) = \exp_p(tv) = \exp_p \circ \phi(t)$, we then have $\exp_{p_*} a = c'(1)$. Also set $w = \tau_v^{-1}(b) \in T_pM$. Let Y be the unique Jacobi field along c with $Y(0) = 0$ and $Y'(0) = w$. By Proposition 10.16, we know that $Y(t) = \exp_{p_*}(t\tau_{tv}w)$. In particular, $Y(1) = \exp_{p_*}(\tau_v w) = \exp_{p_*}(b)$.

From Definition 10.17, we have $\langle\langle a, b\rangle\rangle = \langle \tau_v^{-1}(a), \tau_v^{-1}(b)\rangle = \langle v, w\rangle$. Hence the Gauss Lemma is proved if we show that $\langle v, w\rangle = \langle c'(1), Y(1)\rangle$. But by Lemma 10.9, the function $f(t) = \langle c'(t), Y(t)\rangle = \alpha t + \beta$ for some constants $\alpha, \beta \in \mathbb{R}$. Since $Y(0) = 0$, we have $\beta = 0$ and $f(t) = tf'(0) = t\langle c'(0), Y'(0)\rangle = t\langle v, w\rangle$. In particular, $\langle c'(1), Y(1)\rangle = f(1) = \langle v, w\rangle$ as required. □

The Gauss Lemma has the following geometric consequences. The proofs of these corollaries, which are given along the lines of Gromoll, Klingenberg, and Meyer (1975, pp. 137–138) in Bölts (1977, pp. 75–77), will be omitted. The use of the Gauss Lemma here replaces the use of a synchronous coordinate system in Penrose (1972, p. 53).

Corollary 10.19. *Let U be a convex normal neighborhood in M, and let $c : [0,1] \to U$ be a future directed timelike geodesic segment from $p = c(0)$ to $q = c(1)$ in U. Then if $\beta : [0,1] \to U$ is any future directed timelike piecewise smooth curve from p to q, we have $L(\beta) \leq L(c)$, and $L(\beta) < L(c)$ unless β is just a reparametrization of c.*

The basic idea of the proof is that since U is convex, β and c may be lifted to rays $\widetilde{\beta} : [0,1] \to T_pM$ and $\widetilde{c} : [0,1] \to T_pM$, with $\widetilde{c}(t) = tc'(0)$. Then the Gauss Lemma may be applied to compare $\beta' = \exp_{p_*} \circ \widetilde{\beta}$ and $c' = \exp_{p_*} \circ \widetilde{c}$ and hence the lengths of β and the geodesic c.

An alternative formulation of Corollary 10.19 is also given in Bölts (1977, pp. 75–77) as follows.

Corollary 10.20. *Let $v \in T_pM$ be a timelike tangent vector in the domain of definition of $\exp_p$. Let $\phi : [0,1] \to T_pM$ be the curve $\phi(t) = tv$. Let $\psi : [0,1] \to T_pM$ be a piecewise smooth curve with $\psi(0) = \phi(0)$ and $\psi(1) = \phi(1)$ such that $\exp_p \circ \psi : [0,1] \to M$ is a future directed nonspacelike curve. Then*

$L(\exp\circ\psi) \leq L(\exp\circ\phi)$, and moreover,

$$L(\exp\circ\psi) < L(\exp\circ\phi)$$

provided that there is a $t_0 \in (0,1]$ such that the component b of $\psi'(t_0)$ perpendicular to $\tau_{\psi(t_0)}(\psi(t_0)/\|\psi(t_0)\|)$ satisfies $\exp_{p_*} b \neq 0$.

We are now ready to show that if the timelike geodesic segment c has no conjugate points (recall Definition 10.3), then the length of c is a local maximum.

Proposition 10.21. *Let $c : [a,b] \to M$ be a future directed timelike geodesic segment with no conjugate points to $t = a$ and let $\alpha : [a,b] \times (-\epsilon, \epsilon) \to M$ be any proper piecewise smooth variation of c. Then there exists a constant $\delta > 0$ such that the neighboring curves $\alpha_s : [a,b] \to M$ given by $\alpha_s(t) = \alpha(t,s)$ satisfy $L(\alpha_s) \leq L(c)$ for all s with $|s| < \delta$. Also $L(\alpha_s) < L(c)$ if $0 < |s| < \delta$ unless the curve α_s is a reparametrization of c.*

Proof. Reparametrize α to a variation $\alpha : [0,\beta] \times (-\epsilon,\epsilon) \to M$. By Lemma 10.7, there is an $\epsilon_1 > 0$ such that all the neighboring curves α_s of $\alpha \mid [0,\beta] \times (-\epsilon_1,\epsilon_1)$ are timelike. We may then restrict our attention to $\alpha \mid [0,\beta] \times (-\epsilon_1,\epsilon_1)$.

Set $p = c(0)$ and let $\phi : [0,\beta] \to T_pM$ be the ray $\phi(t) = tc'(0)$. Since c has no conjugate points, $\exp_p$ has maximal rank at $\phi(t) \in T_pM$ for each $t \in [0,\beta]$. Thus by the inverse function theorem, there is a neighborhood of $\phi(t)$ in T_pM which is mapped by $\exp_p$ diffeomorphically onto a neighborhood of $c(t)$ in M. Since $\phi([0,\beta])$ is compact in T_pM, we can find a finite partition $0 = t_0 < t_1 < \cdots < t_k = \beta$ and a neighborhood $U_j \supset \phi([t_j,t_{j+1}])$ in T_pM for each $j = 0,1,\ldots,k-1$ such that $h_j = \exp_p|_{U_j} : U_j \to M$ is a diffeomorphism of U_j onto its image. By continuity, we may then find a constant $\delta_j > 0$ such that $\alpha([t_j,t_{j+1}] \times (-\delta_j,\delta_j)) \subseteq \exp_p(U_j)$ for each $j = 0,1,\ldots,k-1$. Set $\delta = \min\{\delta_1,\delta_2,\ldots,\delta_k\}$.

Then we may define a piecewise smooth map $\Phi : [0,\beta] \times (-\delta,\delta) \to T_pM$ with $\exp_p \circ \Phi = \alpha$ and $\Phi(t,0) = \phi(t)$ for all $t \in [0,\beta]$ as follows. Given $(t,s) \in [0,\beta] \times [-\delta,\delta]$, choose j with $t \in [t_j,t_{j+1}]$, and set $\Phi(t,s) = (h_j)^{-1}(\alpha(t,s))$.

Corollary 10.20 then implies that $L(\alpha_s) = L(\exp_p \circ \Phi) \leq L(\exp_p \circ \phi) = L(c)$ for each s with $|s| \leq \delta$, and equality holds only if α_s is a reparametrization of c. □

With Proposition 10.21 in hand, we may now show that the negative definiteness of the index form on $V_0^{\perp}(c)$ is equivalent to the assumption that c has no points conjugate to $t = a$ along c in $[a, b]$.

Theorem 10.22. *For a given future directed timelike geodesic segment* $c : [a, b] \to M$, *the following are equivalent:*

(1) *The geodesic segment c has no conjugate points to $t = a$ in $(a, b]$.*
(2) *The index form $I : V_0^{\perp}(c) \times V_0^{\perp}(c) \to \mathbb{R}$ is negative definite.*

Proof. (1) $\Rightarrow$ (2) Suppose $Y \neq 0$ in $V_0^{\perp}(c)$ and $I(Y, Y) > 0$. Let $\alpha(t, s) = \exp_{c(t)}(sY(t))$ be the canonical variation associated to Y. Then we have $L'(0) = 0$, $L''(0) = I(Y, Y) > 0$, so that for all $s \neq 0$ sufficiently small, $L(\alpha_s) > L(c)$. But this then contradicts Proposition 10.21. Hence if $Y \neq 0$, then $I(Y, Y) \leq 0$, so that the index form is negative semidefinite.

It remains to show that if $Y \in V_0^{\perp}(c)$ and $I(Y, Y) = 0$, then $Y = 0$. To this end, let $Z \in V_0^{\perp}(c)$ be arbitrary. By Remark 10.2, we have $Y - tZ \in V_0^{\perp}(c)$ for all $t \in \mathbb{R}$. Hence $I(Y - tZ, Y - tZ) \leq 0$ for all $t \in \mathbb{R}$ by the negative semidefiniteness of the index form just established. Since $I(Y - tZ, Y - tZ) = -2tI(Y, Z) + t^2 I(Z, Z)$, it follows that $I(Y, Z) = 0$. As Z was arbitrary, this then implies by Proposition 10.13–(2) that Y is a Jacobi field. Since c has no conjugate points, $Y = 0$.

(2) $\Rightarrow$ (1) Suppose Y is a Jacobi field with $Y(a) = Y(t_1) = 0$, $a < t_1 \leq b$. By Corollary 10.10, $Y \in V^{\perp}(c)$. Extend Y to a nontrivial vector field $Z \in V_0^{\perp}(c)$ by setting

$$Z(t) = \begin{cases} Y(t) & \text{if } a \leq t \leq t_1, \\ 0 & \text{if } t_1 \leq t \leq b. \end{cases}$$

Then $I(Z, Z) = I(Z, Z)_a^{t_1} + I(Z, Z)_{t_1}^b = -\left.\langle Z', Z \rangle\right|_a^{t_1} + 0 = 0.$ □

A consequence of Theorem 10.22 that is crucial to the proof of the Timelike Morse Index Theorem is the following maximality property of Jacobi fields with respect to the index form for timelike geodesic segments without conjugate points. This result is dual to the minimality of Jacobi fields with respect to the index form for geodesics without conjugate points in Riemannian manifolds.

Theorem 10.23 (Maximality of Jacobi Fields). *Let $c : [a,b] \to M$ be a timelike geodesic segment with no conjugate points to $t = a$, and let $J \in V^{\perp}(c)$ be any Jacobi field. Then for any vector field $Y \in V^{\perp}(c)$ with $Y \neq J$, and*

$$Y(a) = J(a) \quad \text{and} \quad Y(b) = J(b), \tag{10.13}$$

we have

$$I(J,J) > I(Y,Y). \tag{10.14}$$

Proof. The vector field $W = J - Y \in V_0^{\perp}(c)$ by (10.13) and $W \neq 0$ as $Y \neq J$ by hypothesis. By Theorem 10.22, we thus have $I(W,W) < 0$. Now calculating $I(W,W)$ we obtain

$$\begin{aligned} I(W,W) &= I(Y,Y) - 2I(J,Y) + I(J,J) \\ &= I(Y,Y) + 2\,\langle J', Y\rangle\big|_a^b - \langle J', J\rangle\big|_a^b . \end{aligned}$$

Since $Y(a) = J(a)$ and $Y(b) = J(b)$, we have $\langle J', Y\rangle\big|_a^b = \langle J', J\rangle\big|_a^b$. Thus

$$\begin{aligned} I(W,W) &= I(Y,Y) + 2\langle J', J\rangle\big|_a^b - \langle J', J\rangle\big|_a^b \\ &= I(Y,Y) + \langle J', J\rangle\big|_a^b = I(Y,Y) - I(J,J). \end{aligned}$$

As $I(W,W) < 0$, this establishes (10.14). □

Now that Theorem 10.23 is obtained, a Morse Index Theorem may be established. First we must define the index of any timelike geodesic $c : [a,b] \to M$. The definition given makes sense because $V_0^{\perp}(c)$ is a vector space.

Definition 10.24. (*Index and Extended Index*) The *index* $\operatorname{Ind}(c)$ and the *extended index* $\operatorname{Ind}_0(c)$ of the timelike geodesic $c : [a,b] \to M$ are defined by

$$\operatorname{Ind}(c) = \operatorname{lub}\{\dim A : A \text{ is a vector subspace of } V_0^{\perp}(c) \text{ and } I \,|\, A \times A \text{ is positive definite}\},$$

and

$$\operatorname{Ind}_0(c) = \operatorname{lub}\{\dim A : A \text{ is a vector subspace of } V_0^\perp(c) \text{ and } I \mid A \times A \text{ is positive semidefinite}\}.$$

Also let $J_t(c)$ denote the $\mathbb{R}$-vector space of smooth Jacobi fields Y along c with $Y(a) = Y(t) = 0$, $a < t \leq b$.

We now relate $\operatorname{Ind}(c)$ to $\operatorname{Ind}_0(c)$ and establish their finiteness in Proposition 10.25. The maximality of Jacobi fields with respect to the index form for timelike geodesics without conjugate points plays a key role in the proof of this proposition. The basic ideas involved in the proof of Proposition 10.25 and Theorem 10.27 are due to Marston Morse (1934).

Proposition 10.25. *Let $c : [a, b] \to M$ be a future directed timelike geodesic segment. Then* $\operatorname{Ind}(c)$ *and* $\operatorname{Ind}_0(c)$ *are finite and*

$$\operatorname{Ind}_0(c) = \operatorname{Ind}(c) + \dim J_b(c). \tag{10.15}$$

Proof. The proof follows the usual method of approximating $V_0^\perp(c)$ by finite-dimensional vector spaces of piecewise smooth Jacobi fields. To this end, choose a finite partition $a = t_0 < t_1 < \cdots < t_k = b$ so that $c \mid [t_i, t_{i+1}]$ has no conjugate points to $t = t_i$ for each i with $0 \leq i \leq k-1$. Let $J\{t_i\}$ denote the subspace of $V_0^\perp(c)$ consisting of all $Y \in V_0^\perp(c)$ such that $Y \mid [t_i, t_{i+1}]$ is a Jacobi field for each i with $0 \leq i \leq k-1$. Since $c \mid [t_i, t_{i+1}]$ has no conjugate points for each i, it follows that $\dim J\{t_i\} = (n-1)(k-1)$.

We now define the approximation $\phi : V_0^\perp(c) \to J\{t_i\}$ of $V_0^\perp(c)$ by $J\{t_i\}$ as follows. For $X \in V_0^\perp(c)$ let $(\phi X) \mid [t_i, t_{i+1}]$ be the unique Jacobi field along $c \mid [t_i, t_{i+1}]$ with $(\phi X)(t_i) = X(t_i)$ and $(\phi X)(t_{i+1}) = X(t_{i+1})$ for each i with $0 \leq i \leq k-1$. Thus X is approximated by a piecewise smooth Jacobi field ϕX such that $(\phi X)(t_i) = X(t_i)$ at each t_i, $0 \leq i \leq k$. This approximation is useful in this context as ϕ is index nondecreasing. More explicitly, $\phi \mid J\{t_i\}$ is just the identity map, so that $I(X, X) = I(\phi X, \phi X)$ if $X \in J\{t_i\}$. On the other hand, if $X \notin J\{t_i\}$ then the inequality

$$I(X, X) < I(\phi X, \phi X) \tag{10.16}$$

may be obtained by applying Theorem 10.23 to each subinterval $[t_i, t_{i+1}]$ and summing.

We establish the following sublemma which shows the finiteness of $\mathrm{Ind}_0(c)$ and $\mathrm{Ind}(c)$ and also allows us to replace $V_0^\perp(c)$ by $J\{t_i\}$ in calculating these indexes.

Sublemma 10.26. *Let* $\mathrm{Ind}_0'(c)$ *and* $\mathrm{Ind}'(c)$ *denote the extended index and index, respectively, of* $I \mid J\{t_i\} \times J\{t_i\}$. *Then*

$$\mathrm{Ind}_0(c) = \mathrm{Ind}_0'(c) \qquad \textit{and} \qquad \mathrm{Ind}(c) = \mathrm{Ind}'(c).$$

Hence $\mathrm{Ind}_0(c)$ *and* $\mathrm{Ind}(c)$ *are finite.*

Proof. First it is easily seen by the uniqueness of the Jacobi field J along $c \mid [t_i, t_{i+1}]$ with $J(t_i) = v$ and $J(t_{i+1}) = w$ that the map $\phi : V_0^\perp(c) \to J\{t_i\}$ is $\mathbb{R}$-linear. That is, if X_1, $X_2 \in V_0^\perp(c)$ and $\alpha, \beta \in \mathbb{R}$, then $\phi(\alpha X_1 + \beta X_2) = \alpha\phi(X_1) + \beta\phi(X_2)$. Thus ϕ maps a vector subspace of $V_0^\perp(c)$ to a vector subspace of $J\{t_i\}$.

In order to establish the sublemma, we first show that if A is any subspace of $V_0^\perp(c)$ on which $I \mid A \times A$ is positive semidefinite, then $\phi|_A : A \to J\{t_i\}$ is injective. Thus suppose $X \in A$ and $\phi(X) = 0$. If $X \in J\{t_i\}$, then $\phi(X) = X$, so that $X = 0$. If $X \notin J\{t_i\}$, then $I(\phi(X), \phi(X)) > I(X, X)$ by (10.16). Thus $\phi X = 0$ implies that $I(X, X) < 0$ which contradicts the assumption that $I \mid A \times A$ is positive semidefinite. Thus if $\phi X = 0$, then $X = 0$.

Now that we know that ϕ is injective on subspaces of $V_0^\perp(c)$ on which I is positive semidefinite, we may prove the sublemma. For if A is a subspace of $V_0^\perp(c)$ on which $I \mid A \times A$ is positive semidefinite, inequality (10.16) implies that the index form of $J\{t_i\}$ is positive semidefinite on the subspace $\phi(A)$ of $J\{t_i\}$. Since ϕ is injective on such a subspace from above, $\dim A = \dim \phi(A)$. Hence $\mathrm{Ind}_0'(c) \geq \mathrm{Ind}_0(c)$. However, since $J\{t_i\}$ is a vector subspace of $V_0^\perp(c)$, we have $\mathrm{Ind}_0'(c) \leq \mathrm{Ind}_0(c)$. Thus $\mathrm{Ind}_0(c) = \mathrm{Ind}_0'(c)$. The same argument shows $\mathrm{Ind}(c) = \mathrm{Ind}(c')$. $\square$

To conclude the proof of Proposition 10.25, we must establish the equality $\mathrm{Ind}_0(c) = \mathrm{Ind}(c) + \dim J_b(c)$. To this end, we choose a second partition

$a = s_0 < s_1 < \cdots < s_m = b$ so that $\{s_1, s_2, \ldots, s_{m-1}\} \cap \{t_1, t_2, \ldots, t_{k-1}\} = \emptyset$ and $c \mid [s_i, s_{i+1}]$ has no conjugate points for each i with $0 \leq i \leq m-1$. Let $J\{s_i\}$ denote the vector subspace of $V_0^\perp(c)$ consisting of all vector fields Y such that $Y \mid [s_i, s_{i+1}]$ is a Jacobi field for each i with $0 \leq i \leq m-1$. Since the two partitions $\{s_i\}$ and $\{t_i\}$ are distinct except for $a = s_0 = t_0$ and $b = s_m = t_k$, it follows that

$$J\{t_i\} \cap J\{s_i\} = J_b(c)$$

where $J_b(c)$ denotes the vector space of all (smooth) Jacobi fields J along c with $J(a) = J(b) = 0$. By Corollary 10.10, we have $J_b(c) \subseteq V_0^\perp(c)$. Also, if $X \in J\{s_i\}$ but $X \notin J_b(c)$, we have by (10.16) that

$$I(X,X) < I(\phi X, \phi X). \tag{10.17}$$

Applying the proof of Sublemma 10.26 to the partition $\{s_i\}$ of $[a,b]$, we may choose a vector subspace B_0' of $J\{s_i\}$ with $I \mid B_0' \times B_0'$ positive semidefinite and with $\operatorname{Ind}_0(c) = \dim B_0'$. Since $\dim B_0' = \operatorname{Ind}_0(c) < \infty$, it follows that $J_b(c)$ is a vector subspace of B_0'. By (10.17) and the proof of Sublemma 10.26, the map

$$\phi\,|_{B_0'} : B_0' \to J\{t_i\}$$

is injective. Thus if we set $B_0 = \phi(B_0')$, we have $\dim B_0 = \dim B_0' = \operatorname{Ind}_0(c)$. Since B_0 is a finite-dimensional vector space and $J_b(c)$ is a vector subspace, we may find a vector subspace B of B_0 such that $B_0 = B \oplus J_b(c)$, where $\oplus$ denotes the direct sum of vector spaces.

We claim now that $I \mid B \times B$ is positive definite. By construction, we know that $I \mid B_0' \times B_0'$ is positive semidefinite. Also if $0 \neq Z \in B$ and we represent $Z = \phi(X)$ with $X \in B_0'$, then $X \notin J_b(c)$. (For if $X \in J_b(c)$, we have $\phi(X) = X$ so that $Z = \phi(X) = X \in J_b(c)$ also, which is impossible since $B \cap J_b(c) = \{0\}$ by construction.) Hence $I(Z,Z) = I(\phi X, \phi X) > I(X,X)$, the last inequality by (10.17). Thus as $I \mid B_0' \times B_0'$ is positive semidefinite, we obtain $I(Z,Z) > I(X,X) \geq 0$ so that $I(Z,Z) > 0$. This shows that $I \mid B \times B$ is positive definite. With the notation of Sublemma 10.26, we then have $\operatorname{Ind}'(c) \geq \dim B$.

From the direct sum decomposition $B_0 = B \oplus J_b(c)$, we obtain the equality

$$\mathrm{Ind}_0(c) = \dim B_0 = \dim B + \dim J_b(c),$$

and we also know that $\mathrm{Ind}'(c) \geq \dim B$. Thus the proof of Proposition 10.25 will be complete if we show that $\mathrm{Ind}'(c) \leq \dim B$.

To establish this inequality, suppose $B' \subseteq J\{t_i\}$ is a vector subspace with $I \,|\, B' \times B'$ positive definite and $\dim B' = \mathrm{Ind}'(c)$. Suppose that $\dim B' > \dim B$. Then I is positive semidefinite on $B' \oplus J_b(c)$ so that $\dim(B' \oplus J_b(c)) \leq \mathrm{Ind}_0(c)$. On the other hand, since $\dim B' > \dim B$ we obtain

$$\dim(B' \oplus J_b(c)) > \dim(B \oplus J_b(c)) = \mathrm{Ind}_0(c).$$

This contradiction shows that $\dim B' \leq \dim B$, whence $\mathrm{Ind}'(c) \leq \dim B$ as required. □

We are now in a position to prove a Morse Index Theorem for timelike geodesic segments. The proof we give here is modeled on the proof for Riemannian spaces given in Gromoll, Klingenberg, and Meyer (1975, pp. 150–152).

Theorem 10.27 (Timelike Morse Index Theorem). *Let $c : [a,b] \to M$ be a timelike geodesic segment, and for each $t \in [a,b]$ let $J_t(c)$ denote the $\mathbb{R}$-vector space of smooth Jacobi fields Y along c with $Y(a) = Y(t) = 0$. Then c has only finitely many conjugate points, and the index* $\mathrm{Ind}(c)$ *and extended index* $\mathrm{Ind}_0(c)$ *of the index form $I : V_0^{\perp}(c) \times V_0^{\perp}(c) \to \mathbb{R}$ are given by the formulas*

(10.18)
$$\mathrm{Ind}(c) = \sum_{t \in (a,b)} \dim J_t(c)$$

and

(10.19)
$$\mathrm{Ind}_0(c) = \sum_{t \in (a,b]} \dim J_t(c),$$

respectively.

Proof. We first show that $\sum_{t \in (a,b]} \dim J_t(c)$ is a finite sum. We know that $\dim J_t(c) \geq 1$ if and only if $c(t)$ is a conjugate point of $t = a$ along c. We may also define embeddings

$$i : J_t(c) \to V_0^{\perp}(c)$$

for each $t \in [a, b]$ by

$$i(Y)(s) = \begin{cases} Y(s) & \text{for } a \le s \le t, \\ 0 & \text{for } t \le s \le b. \end{cases}$$

Evidently $\dim J_t(c) = \dim i(J_t(c))$ for any $t \in (a, b]$.

Recall that $\mathrm{Ind}_0(c)$ is finite by Proposition 10.25. Thus to show that c has only finitely many conjugate points, it suffices to prove that if $\{t_1, t_2, \ldots, t_k\}$ is any set of pairwise distinct conjugate points to $t = a$ along c, then $k \le \mathrm{Ind}_0(c)$. To this end, set $A_j = i(J_{t_j}(c))$ for each $j = 1, 2, \ldots, k$ and $A = A_1 \oplus \cdots \oplus A_k$. Then A is a vector subspace of $V_0^\perp(c)$, and decomposing $Z \in A$ as $Z = \sum_{j=1}^k \lambda_j Z_j$ with $\lambda_j \in \mathbb{R}$, $Z_j \in A_j$, we obtain $I(Z, Z) = \sum_{j,l} \lambda_j \lambda_l I(Z_j, Z_l)$. But if $t_j \le t_l$, we obtain $I(Z_j, Z_l) = I(Z_j, Z_l)_a^{t_l} + I(Z_j, Z_l)_{t_l}^b = -\langle Z_l', Z_j \rangle|_a^{t_l} + I(0, Z_l)_{t_l}^b = 0$ using (10.2), $Z_l \in A_l$, and $Z_j(a) = Z_j(t_l) = 0$. Hence $I(Z, Z) = 0$ from the symmetry of the index form. Thus $I \,|\, A \times A$ is positive semidefinite. Hence

$$k \le \dim A = \dim A_1 + \cdots + \dim A_k \le \mathrm{Ind}_0(c)$$

as required. Therefore $c : [a, b] \to M$ has only finitely many conjugate points in $(a, b]$, which we denote by $t_1 < t_2 < \cdots < t_r$. Except for $t \in \{t_1, t_2, \ldots, t_r\}$, we have $\dim J_t(c) = 0$, and thus $\sum_{t \in (a,b]} \dim J_t(c)$ is a finite sum.

Since $\mathrm{Ind}_0(c) = \mathrm{Ind}(c) + \dim J_b(c)$ from Proposition 10.25, it is enough to establish the equality

$$\mathrm{Ind}_0(c) = \sum_{t \in (a,b]} \dim J_t(c) \tag{10.20}$$

to prove Theorem 10.27. Let $\mathbb{Z}$ denote the integers with the discrete topology and define $f, f_0 : (a, b] \to \mathbb{Z}$ by $f(t) = \mathrm{Ind}(c \,|\, [a, t])$ and $f_0(t) = \mathrm{Ind}_0(c \,|\, [a, t])$. We now show that (10.20) holds if we establish the left continuity of f and the right continuity of f_0. Here we mean that $\lim_{t_n \uparrow t} f(t_n) = f(t)$ and $\lim_{t_n \downarrow t} f_0(t_n) = f_0(t)$. Using (10.15) it follows that $f(t) - f_0(t) = -\dim J_t(c) = 0$ if $t \notin \{t_1, t_2, \ldots, t_r\}$. Assuming we have shown f is left continuous and f_0 is right continuous, we also have $f(t_{j+1}) = f_0(t_j)$ for each $j = 1, 2, \ldots, r - 1$.

Thus

$$\sum_{t\in(a,b]} \dim J_t(c) = \sum_{t\in(a,b]} [f_0(t) - f(t)]$$
$$= \sum_{j=1}^{r} [f_0(t_j) - f(t_j)] = f_0(t_r) - f(t_1).$$

By Theorem 10.22 the index form is negative definite if c has no conjugate points, so that $f(t) = 0$ for all $t < t_1$. Hence $f(t_1) = 0$ since f is left continuous. Thus $\sum_{t\in(a,b]} \dim J_t(c) = f_0(t_r)$. Since f_0 is constant on $[t_r, b]$, we have $f_0(t_r) = f_0(b)$. Hence

$$\sum_{t\in(a,b]} \dim J_t(c) = f_0(t_r) = f_0(b) = \mathrm{Ind}_0(c\,|\,[a,b])$$

which establishes (10.20).

We have thus reduced the proof of the Morse Index Theorem to showing that f is left continuous and f_0 is right continuous. First note that f and f_0 are nondecreasing (i.e., $f(t) \geq f(s)$ if $t \geq s$). For suppose we fix $t_1, t_2 \in (a,b]$ with $t_1 \leq t_2$. Let $c_1 = c\,|\,[a,t_1]$ and $c_2 = c\,|\,[a,t_2]$. We then have an $\mathbb{R}$-linear embedding $i : V_0^\perp(c_1) \to V_0^\perp(c_2)$ given by

$$i(Y)(t) = \begin{cases} Y(t) & \text{for } a \leq t \leq t_1, \\ 0 & \text{for } t_1 \leq t \leq t_2. \end{cases}$$

This map has the property that $I(Y,Y) = I(i(Y), i(Y))$, where the indexes are calculated with respect to c_1 and c_2, respectively. Thus if $A \subseteq V_0^\perp(c_1)$ is a vector subspace on which the index form of c_1 is positive (semi) definite, then $i(A)$ is a vector subspace of $V_0^\perp(c_2)$ on which the index form of c_2 is positive (semi) definite and $\dim A = \dim i(A)$. Thus $f(t_1) = \mathrm{Ind}(c\,|\,[0,t_1]) \leq \mathrm{Ind}(c\,|\,[0,t_2]) = f(t_2)$ and similarly $f_0(t_1) \leq f_0(t_2)$. Thus f_0 and f are nondecreasing.

To obtain the continuity properties of f and f_0, we fix an arbitrary $\widetilde{t} \in (a,b]$ and study the continuity of f and f_0 at $\widetilde{t}$ using the same approximation techniques as in the proof of Sublemma 10.26. Since $c([a,b])$ is a compact subset of M, there is a constant $\delta > 0$ such that for any $s_1, s_2 \in [a,b]$ with $|s_1 - s_2| < \delta$ and $s_1 \leq s_2$, the geodesic segment $c\,|\,[s_1,s_2]$ has no conjugate points. Choose a partition $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = \widetilde{t}$ (unrelated to the above enumeration of conjugate points) such that $|t_i - t_{i+1}| < \delta$ for each $i = 0, 1, 2, \ldots, k-1$. Let $J \subseteq [a, b]$ be an open interval containing $\widetilde{t}$ with $|t - t_{k-1}| < \delta$ for all $t \in J$. For each $t \in J$, let $\widetilde{J}(c_t)$ denote the finite-dimensional $\mathbb{R}$-vector subspace of $V_0^{\perp}(c\,|\,[a,t])$ consisting of all vector fields $Y \in V_0^{\perp}(c\,|\,[a,t])$ such that $Y\,|\,[t_j, t_{j+1}]$ for each $j = 0, 1, \ldots, k-2$ and $Y\,|\,[t_{k-1}, t]$ are Jacobi fields. By Sublemma 10.26, $f(t)$ is the index of I restricted to $\widetilde{J}(c_t) \times \widetilde{J}(c_t)$ and $f_0(t)$ is the extended index of I restricted to $\widetilde{J}(c_t) \times \widetilde{J}(c_t)$.

Now let $E = N(c(t_1)) \times N(c(t_2)) \times \cdots \times N(c(t_{k-1}))$. The set E is closed since each $N(c(t_i)) = \{v \in T_{c(t_i)}M : \langle v, c'(t_i)\rangle = 0\}$ is a closed set of spacelike tangent vectors. We may define a Euclidean product metric $\langle\langle\ ,\ \rangle\rangle : E \times E \to \mathbb{R}$ by $\langle\langle v, w\rangle\rangle = \sum_{i=1}^{k-1}\langle v_i, w_i\rangle$ for $v = (v_1, v_2, \ldots, v_{k-1})$, $w = (w_1, w_2, \ldots, w_{k-1}) \in E$. Then by Remark 10.2, $S = \{v \in E : \|v\| = 1\}$ is compact.

If $Y \in \widetilde{J}(c_t)$, then $Y(a) = Y(t) = 0$ by definition. Also since $c\,|\,[t_i, t_{i+1}]$ has no conjugate points, for any $v \in N(c(t_i))$ and $w \in N(c(t_{i+1}))$ there is a unique Jacobi field Y along c with $Y(t_i) = v$ and $Y(t_{i+1}) = w$. Since $\langle Y, c'\rangle|_t$ is an affine function of t and $\langle v, c'(t_i)\rangle = \langle w, c'(t_{i+1})\rangle = 0$, it follows that $\langle Y, c'\rangle|_t = 0$ for all t. Hence the map

$$\phi_t : \widetilde{J}(c_t) \to E$$

defined for $t \in J$ by

$$\phi_t(Y) = (Y(t_1), Y(t_2), \ldots, Y(t_{k-1}))$$

is an isomorphism. For each $t \in J$ we also have a quadratic form $Q_t : E \times E \to \mathbb{R}$ given by $Q_t(u, v) = I(\phi_t^{-1}(u), \phi_t^{-1}(v))$. Sublemma 10.26 implies that $f_0(t)$ is the extended index of the quadratic form Q_t on $E \times E$ and $f(t)$ is the index of Q_t on $E \times E$ for each $t \in J$.

Each Q_t may then be used to define a map $Q : E \times E \times J \to \mathbb{R}$ given by $Q(u, v, t) = Q_t(u, v)$. We want to show that Q is continuous in order to prove that f and f_0 have the desired continuity properties. To this end, let

$B = \{Y \,|\, [a, t_{k-1}] : Y \in \widetilde{J}(c_t)\}$. Then B is isomorphic to E via the mapping $\phi : B \to E$ given by

$$\phi(Y) = (Y(t_1), Y(t_2), \ldots, Y(t_{k-1}))\,.$$

Thus

$$\begin{aligned} Q(u,v,t) &= I(\phi_t^{-1}(u), \phi_t^{-1}(v)) \\ &= I(\phi^{-1}u, \phi^{-1}v) - \langle X_{u,t}, Y'_{v,t}\rangle\big|_{t_{k-1}}^{t} \end{aligned}$$

where $X_{u,t}$ and $Y_{v,t}$ are the Jacobi fields along c given by

$$X_{u,t} = \phi_t^{-1}(u) \,|\, [t_{k-1}, t]$$

and

$$Y_{v,t} = \phi_t^{-1}(v) \,|\, [t_{k-1}, t].$$

Since the map $(u, v) \to I(\phi^{-1}(u), \phi^{-1}(v))$ from $E \times E \to \mathbb{R}$ is a bilinear form, the map $(u, v, t) \to I(\phi^{-1}(u), \phi^{-1}(v))$ is certainly continuous. By Proposition 10.16, the map $(u, v, t) \to \langle X_{u,t}, Y'_{v,t}\rangle\big|_{t_{k-1}}^{t}$ is continuous. This establishes the continuity of $Q : E \times E \times J \to \mathbb{R}$.

We are finally ready to show that f_0 is right continuous and f is left continuous at the arbitrary $\widetilde{t} \in (a, b]$. Since Sublemma 10.26 implies that f is finite-valued, we may choose a subspace A of E with $\dim A = f(\widetilde{t})$ and $Q(u, u, \widetilde{t}) > 0$ for all $u \in A$, $u \neq 0$. Since $Q : E \times E \times J \to \mathbb{R}$ is continuous and $S_1 = S \cap A = \{u \in A : \|u\| = 1\}$ is compact, there is a neighborhood J_0 of $\widetilde{t}$ in J such that $Q(u, u, t) > 0$ for all $t \in J_0$ and $u \in S_1$. Hence $Q_t \,|\, A \times A$ is positive definite for all $t \in J_0$. Thus $f(t) \geq f(\widetilde{t})$ for all $t \in J_0$. Since f is nondecreasing, it follows that $f(t) = f(\widetilde{t})$ for all $t \in J_0$ with $t \leq \widetilde{t}$. Hence f is left continuous.

It remains to show the right continuity of f_0 at $\widetilde{t}$. Let $\{s_n\}$ be a sequence in J with $s_n > t$ for all n and $s_n \to \widetilde{t}$. Since f_0 is nondecreasing and integer-valued, we may suppose that $f_0(s_n) = k$ for all n. By the monotonicity of f_0, we then have $f_0(\widetilde{t}) \leq k$. It thus remains to show that $f_0(\widetilde{t}) \geq k$. To accomplish this, choose for each n a k-dimensional subspace A_n of E such that $Q_{s_n} \mid A_n \times A_n$ is positive semidefinite. Let $\{a_1(n), a_2(n), \ldots, a_k(n)\}$ be an orthonormal basis of A_n for each n. Thus $a_1(n), a_2(n), \ldots, a_k(n)$ are contained in the compact subset S of E. By the compactness of S, we may assume $a_j(n) \to a_j \in S$ for each j. By continuity of the inner product, it follows that the vectors $\{a_1, a_2, \ldots, a_k\}$ form an orthonormal subset of S. Let $A = \operatorname{span}\{a_1, a_2, \ldots, a_k\}$, which is thus a k-dimensional subspace of E. Given $u \in A$, we may write $u = \sum_{j=1}^{k} \lambda_j a_j$. Let $u(n) = \sum_{j=1}^{k} \lambda_j a_j(n)$. Obviously $u(n) \to u$ as $n \to \infty$. Thus using the continuity of $Q : E \times E \times J \to \mathbb{R}$, we obtain

$$Q_{\widetilde{t}}(u, u) = \lim_{n \to \infty} Q(u(n), u(n), s_n) \geq 0$$

since $Q_{s_n} \mid A_n \times A_n$ is positive semidefinite for each n. Hence $Q_{\widetilde{t}} \mid A \times A$ is positive semidefinite. Thus $f_0(\widetilde{t}) \geq \dim A = k$ as required. □

We conclude this section with an application of the Timelike Morse Index Theorem to the structure of the cut locus of future one-connected, globally hyperbolic space–times [cf. Beem and Ehrlich (1979c, Section 8)].

Definition 10.28. (*Future One-Connected Space–Time*) A space–time (M, g) is said to be *future one-connected* if for all $p, q \in M$, any two future directed timelike curves from p to q are homotopic through smooth future directed timelike curves with fixed endpoints p and q.

This concept, a Lorentzian analogue for simple connectivity, has been studied in Avez (1963), Smith (1960a), and Flaherty (1975a, p. 395). The vanishing of the Lorentzian fundamental group implies that (M, g) is future one-connected. However, the simple connectivity of M as a topological space does *not* imply that (M, g) is future one-connected, as the following example of Geroch shows. Consider $\mathbb{R}^3$ with coordinates (x, y, t) and the Lorentzian metric $ds^2 = -dt^2 + dx^2 + dy^2$. Let $T = \{(x, 0, 0) \in \mathbb{R}^3 : x \geq 0\}$ and set $M = \mathbb{R}^3 - T$ with the induced Lorentzian metric from $(\mathbb{R}^3, ds^2)$. Then M is simply connected. On the other hand, let $p = (2, 0, -1)$ and $q = (2, 0, 1)$. Then p and q may be joined by future directed timelike curves γ_1 and γ_2 lying on opposite sides of the positive x axis (cf. Figure 10.1). But γ_1 and γ_2 are not homotopic through future directed timelike curves starting at p and ending at q since such

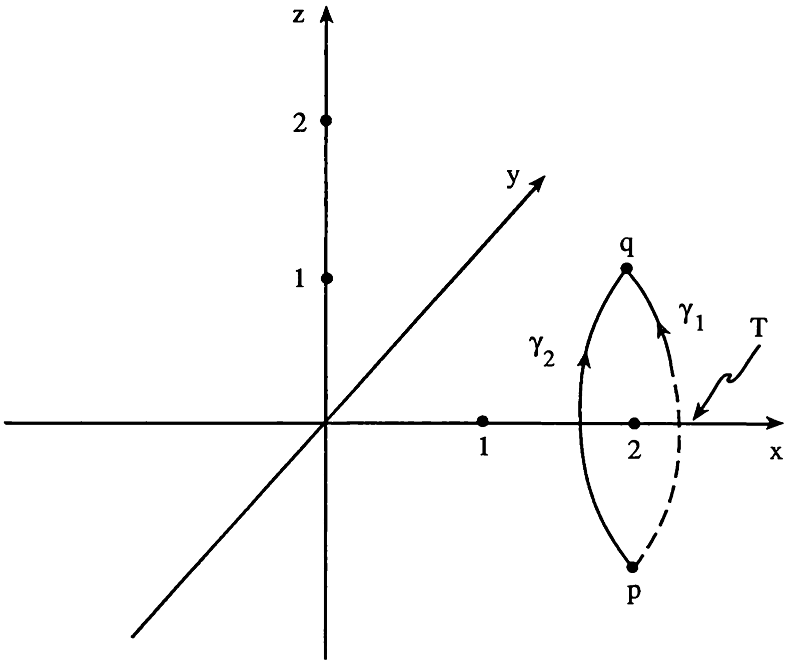


FIGURE 10.1. A space–time $M = \mathbb{R}^3 - T$ which is simply connected but not future one-connected is shown. The future directed timelike curves γ_1 and γ_2 from p to q are not homotopic through timelike curves with endpoints p and q.

a homotopy would have to go around the point $(0, 0, 0)$, which would introduce spacelike curves.

Note that if (M, g) is future one-connected, then the path space of smooth timelike curves from p to q is connected. Thus Lemma 4.11–(2) of Cheeger and Ebin (1975, p. 85) and the standard path space Morse theory [cf. Everson and Talbot (1976), Uhlenbeck (1975), Woodhouse (1976)] imply the following proposition.

Proposition 10.29. *Let (M, g) be future one-connected and globally hyperbolic. Fix $p \in M$ and suppose that every future directed timelike geodesic segment starting at p has index zero or index greater than or equal to two. Given $q \in I^+(p)$ such that q is not conjugate to p along any future directed timelike geodesic from p to q, there is exactly one future directed timelike geodesic from p to q of index zero, namely, the unique maximal geodesic from p to q.*

We are now ready to prove the Lorentzian analogue for globally hyperbolic, future one-connected space–times of a theorem of Cheeger and Ebin (1975, Theorem 5.11) on the cut locus of a complete Riemannian manifold which generalized a theorem of Crittenden (1962) for simply connected Lie groups with bi-invariant Riemannian metrics. The global hyperbolicity is used in Theorem 10.30 to guarantee the existence of maximal geodesic segments joining chronologically related points. Here the *order* of a conjugate point $t = t_0$ to p along the timelike geodesic γ with $\gamma(0) = p$ is defined as $\dim J_{t_0}(\gamma)$ [recall Definition 10.24].

Theorem 10.30. *Let (M, g) be future one-connected and globally hyperbolic. Suppose that for $p \in M$, the first future conjugate point of p along every timelike geodesic γ with $\gamma(0) = p$ is of order two or greater. Then the future timelike cut locus of p and the locus of first future timelike conjugate points of p coincide. Thus all future timelike geodesics from p maximize up to the first future conjugate point.*

Proof. All geodesics will be unit speed and future timelike during the course of this proof. It suffices to show that if $\gamma : [0, t] \to M$ with $\gamma(0) = p$ has index zero and $\gamma(t)$ is not conjugate to p along $\gamma \,|\, [0, t]$, then γ is maximal. Since the set of singular points of $\exp_p$ is closed, it follows from Theorem 10.27 that there exist $\epsilon_1, \epsilon_2 > 0$ such that if $\sphericalangle(v, \gamma'(0)) < \epsilon_1$ and $\epsilon < \epsilon_2$, then the future timelike geodesic $\sigma : [0, t + \epsilon] \to M$ with $\sigma'(0) = v$ is of index zero. Here $\sphericalangle$ denotes angle measure using an auxiliary Riemannian metric.

By Sard's Theorem, we may find a sequence of points $\{p_i\} \subseteq I^+(p)$ with $p_i \to \gamma(t)$ such that every timelike geodesic segment from p to p_i has nonconjugate endpoints. Thus, by hypothesis, every such segment has index zero or index two or greater. By Proposition 10.29, there is a unique timelike maximal geodesic segment γ_i from p to p_i of index zero.

Since $(\exp_p)_*$ is nonsingular at $t\gamma'(0)$, for i sufficiently large there are geodesic segments $\overline{\gamma}_i$ from p to p_i with $\overline{\gamma}_i \to \gamma$. Since $\sphericalangle(\overline{\gamma}_i{}'(0), \gamma'(0)) \to 0$, these segments have index zero for large i. It follows that for i sufficiently large, $\overline{\gamma}_i = \gamma_i$ and hence $\overline{\gamma}_i$ is maximal. Thus γ is maximal as a limit of maximal geodesics. □

10.2 The Timelike Path Space of a Globally Hyperbolic Space–time

In this section we discuss the Morse theory of the path space of future directed timelike curves joining two chronologically related points in a globally hyperbolic space–time following Uhlenbeck (1975) [cf. also Woodhouse (1976)]. Both of these treatments are modeled on Milnor's exposition (1963, pp. 88–92) of the Morse theory for the path space of a complete Riemannian manifold in which the full path space is approximated by piecewise smooth geodesics. A different approach has been given to the Morse theory of nonspacelike curves in globally hyperbolic space–time by Everson and Talbot (1976, 1978). They use a result of Clarke (1970) that any four-dimensional globally hyperbolic space–time may be isometrically embedded in a high-dimensional Minkowski space–time to give a Hilbert manifold structure to a subclass of timelike curves in M.

We now turn to Uhlenbeck's treatment of the Morse theory of the path space of piecewise smooth timelike curves joining points $p \ll q$ of any globally hyperbolic space–time (M, g) of dimension $n \geq 2$.

Definition 10.31. (*The Timelike Path Space $C_{(p,q)}$*) Given $p, q \in (M, g)$ with $p \ll q$, let $C_{(p,q)}$ denote the space of future directed piecewise smooth timelike curves from p to q, where two curves which differ by a parametrization are identified.

While $C_{(p,q)}$ is not a manifold modeled on a Banach space, it does possess tangent spaces consisting of piecewise smooth vector fields along the given piecewise smooth curve assumed to be parametrized by arc length. Functionals

$F : C_{(p,q)} \to \mathbb{R}$ may then be considered from the point of view of the calculus of variations. Thus a *critical point* of F is a point at which all first variations vanish, and a *critical value* is the image under F of a critical point. The functional F on $C_{(p,q)}$ is said to be a *homotopic Morse function* if for any $b > a$ which is not a critical value of F, the topological space $F^{-1}(-\infty, b)$ is homotopically equivalent to the space $F^{-1}(-\infty, a)$ with cells adjoined, where one cell of dimension equal to the index of the corresponding critical point is adjoined for each critical point of F with critical value in (a, b).

We will show in this section that the Lorentzian arc length functional is a homotopic Morse function for $C_{(p,q)}$ provided that p and q are nonconjugate along any nonspacelike geodesic. This result is analogous to Morse's result [cf. Milnor (1963, Theorems 16.3 and 17.3)] for complete Riemannian manifolds. Namely, if p and q are any two distinct points not conjugate along any geodesic, then the space $\Omega_{(p,q)}$ of piecewise smooth curves from p to q has the homotopy type of a countable CW-complex with a cell of dimension λ for each geodesic from p to q of index λ.

Given that p and q must be nonconjugate along any geodesic for L to be a Morse function, it is of interest to know that such pairs of points exist. As for Riemannian spaces, conjugate points in an arbitrary Lorentzian manifold may be viewed as singularities of the differential of the exponential mapping. Hence Sard's Theorem [cf. Hirsch (1976, p. 69)] may be applied. Here a subset X of a manifold is said to have *measure zero* if for every chart (U, ϕ), the set $\phi(U \cap X) \subseteq \mathbb{R}^n$ has Lebesgue measure zero in $\mathbb{R}^n$, $n = \dim M$. Also a subset of a manifold is said to be *residual* if it contains the intersection of countably many dense open sets. A residual subset of a complete metric space is dense by the Baire Category Theorem [cf. Kelley (1955, p. 200)]. Sard's Theorem then implies the following result.

Theorem 10.32. *Let (M, g) be a globally hyperbolic space–time, and let $p \in M$ be arbitrary. Then the set of points of M conjugate to p along some geodesic has measure zero. Thus for a residual set of $q \in M$, p and q are nonconjugate along all geodesics between them.*

Recall the following properties of the Lorentzian arc length functional $L : C_{(p,q)} \to \mathbb{R}$. First, from the calculus of variations, the critical points of L on $C_{(p,q)}$ are exactly the future directed timelike geodesic segments from p to q with parameter proportional to arc length. Second, the Timelike Morse Index Theorem (Theorem 10.27) implies that the index of a critical point of L is just its index as a geodesic, i.e., the number of conjugate points to p along the geodesic, counting multiplicities, from p up to but not including q.

In order to show that L is a homotopic Morse function, it is necessary to approximate $C_{(p,q)}$ by a subset $M_{(p,q)}$ such that there exists a retraction of $C_{(p,q)}$ onto $M_{(p,q)}$ which increases the length functional L. This step corresponds to the finite-dimensional approximation of the loop space in Riemannian Morse theory. The corresponding Lorentzian approximation makes crucial use of the global hyperbolicity of (M, g), as will be apparent from Lemma 10.34 below. First it is useful to make the following definition.

Definition 10.33. (*Timelike Chain*) Let $p, q \in (M, g)$ with $p \ll q$. A finite collection $\{x_1, x_2, \ldots, x_j\}$ of points of M is said to be a *timelike chain from p to q* if $p \ll x_1 \ll x_2 \ll \cdots \ll x_j \ll q$.

The next lemma follows from the existence of convex normal neighborhoods (cf. Section 3.1), the compactness of $J^+(p) \cap J^-(q)$ in a globally hyperbolic space–time, and Theorem 6.1.

Lemma 10.34. *Let (M, g) be a globally hyperbolic space–time with a fixed smooth globally hyperbolic time function $f : M \to \mathbb{R}$. Let $p, q \in M$ with $p \ll q$ be given. Then there exists $\{t_1, t_2, \ldots, t_k\}$ with $f(p) < t_1 < t_2 \cdots < t_k < f(q)$ satisfying the following properties:*

(1) *If $x \in f^{-1}(f(p), t_1]$ and $p \leqslant x$, then there is a unique maximal future directed nonspacelike geodesic segment from p to x;*

(2) *For each i with $1 \leq i \leq k-1$, if $x \in f^{-1}(t_i)$ and $y \in f^{-1}(t_i, t_{i+1}]$ with $p \leqslant x \leqslant y \leqslant q$, then there is a unique maximal future directed nonspacelike geodesic segment from x to y;*

(3) *If $y \in f^{-1}[t_k, f(q))$ and $y \leqslant q$, then there is a unique maximal future directed nonspacelike geodesic segment from y to q;*

(4) *In particular, if* $\{x_1, x_2, \ldots, x_k\}$ *is any timelike chain from* p *to* q *with* $x_i \in f^{-1}(t_i)$ *for each* $i = 1, 2, \ldots, k$*, then there is a unique maximal future directed timelike geodesic segment from* p *to* x_1*,* x_k *to* q*, and* x_i *to* x_{i+1} *for each* i *with* $1 \leq i \leq k-1$.

Since $p, q \in M$ with $p \ll q$ are given, we may now fix $\{t_1, t_2, \ldots, t_k\}$ satisfying the conditions of Lemma 10.34. Denote by S_i the Cauchy surface $S_i = f^{-1}(t_i)$ for each $i = 1, 2, \ldots, k$. We may now define a space $M_{(p,q)}$ of broken timelike geodesics to approximate $C_{(p,q)}$ from the point of view of the arc length functional.

Definition 10.35. (*Timelike Curve Space* $M_{(p,q)}$) Let $M_{(p,q)}$ be the space of all continuous curves $\gamma : [0,1] \to M$ such that $\gamma(0) = p$, $\gamma(1) = q$, and $\gamma(i/(k+1)) \in S_i$ for each $i = 1, 2, \ldots, k$ and such that the restricted curve $\gamma \mid [\, i/(k+1), (i+1)/(k+1)\,]$ is a future directed timelike geodesic for each $i = 0, 1, \ldots, k$.

Since each $\gamma \mid [\, i/(k+1), (i+1)/(k+1)\,]$ must be the unique maximal segment joining its endpoints by Lemma 10.34, it follows that

$$(10.21)\quad M_{(p,q)} = \{(x_1, x_2, \ldots, x_k) : x_i \in S_i \text{ for } 1 \leq i \leq k, \\ p \ll x_1,\ \ x_i \ll x_{i+1} \text{ for each } 1 \leq i \leq k-1, \text{ and } x_k \ll q\}.$$

Since chronological future and past sets are open, $M_{(p,q)}$ viewed as in (10.21) is an open submanifold of $S_1 \times S_2 \times \cdots \times S_k$. Let $\pi_i : M_{(p,q)} \to S_i$ be the projection map given by $\pi_i(x_1, x_2, \ldots, x_k) = x_i$ for each $i = 1, 2, \ldots, k$. Since $I^+(p) \cap I^-(q)$ has compact closure by the global hyperbolicity of M, and $\pi_i(M_{(p,q)}) \subseteq I^+(p) \cap I^-(q)$, it follows that $\pi_i(M_{(p,q)})$ has compact closure in S_i for each $i = 1, 2, \ldots, k$.

A length-nondecreasing retraction of $C_{(p,q)}$ onto $M_{(p,q)}$ may now be established along the lines of Riemannian Morse theory [cf. Milnor (1963, p. 91)].

Proposition 10.36. *There exists a retraction* Q_λ*,* $0 \leq \lambda \leq 1$*, of* $C_{(p,q)}$ *onto* $M_{(p,q)}$ *which is Lorentzian arc length nondecreasing.*

Proof. Let $\gamma \in C_{(p,q)}$ be arbitrary. We may suppose that γ is parametrized so that $f(\gamma(t)) = t$ where $f : M \to \mathbb{R}$ is the globally hyperbolic time function fixed in Lemma 10.34. Thus $\gamma : [f(p), f(q)] \to M$. For each $\lambda \in [0,1]$, define $Q_\lambda(\gamma) : [f(p), f(q)] \to M$ as follows. Set $\beta = (1-\lambda)f(p) + \lambda f(q)$, and put $t_0 = f(p)$ and $t_{k+1} = f(q)$. If β satisfies $t_i < \beta \leq t_{i+1}$ for i with $0 \leq i \leq k$, let $Q_\lambda(\gamma)(t)$ be the unique broken timelike geodesic joining the point p to $\gamma(t_1)$, $\gamma(t_1)$ to $\gamma(t_2)$, ... , $\gamma(t_{i-1})$ to $\gamma(t_i)$, and $\gamma(t_i)$ to $\gamma(\beta)$ successively for $t \leq \beta$, and let $Q_\lambda(\gamma)(t) = \gamma(t)$ for $t \geq \beta$ (cf. Figure 10.2). It is immediate from the definition that $Q_0(\gamma) = \gamma$ and that $Q_1(\gamma) \in M_{(p,q)}$. Since $Q_\lambda(\gamma)[\gamma(t_i), \gamma(\beta)]$ is maximal from $\gamma(t_i)$ to $\gamma(\beta)$ by Lemma 10.34, we have

$$L(Q_\lambda(\gamma)[\gamma(t_i), \gamma(\beta)]) = d(\gamma(t_i), \gamma(\beta)) \geq L(\gamma \,|\, [t_i, \beta])$$

where d denotes the Lorentzian distance function of (M, g) and we have used Notational Convention 8.4. Similarly, since $Q_\lambda(\gamma)$ is the unique maximal geodesic segment from $\gamma(t_j)$ to $\gamma(t_{j+1})$ for each j with $0 \leq j \leq i-1$, we have $L(Q_\lambda(\gamma)[\gamma(t_j), \gamma(t_{j+1})]) \geq L(\gamma \,|\, [t_j, t_{j+1}])$ for each j with $0 \leq j \leq i-1$. Summing, we obtain $L(Q_\lambda(\gamma)[p, \gamma(\beta)]) \geq L(\gamma \,|\, [f(p), \beta])$. Since $Q_\lambda(\gamma)(t) = \gamma(t)$ for $t \geq \beta$, we thus have $L(Q_\lambda(\gamma)) \geq L(\gamma)$. Moreover, it is clear from the above argument that $L(Q_\lambda(\gamma)) = L(\gamma)$ for all λ if and only if γ is a broken timelike geodesic from p to q with breaks possible only at the S_i's. In particular, $Q_\lambda \,|\, M_{(p,q)} = \mathrm{Id}$ for all $\lambda \in [0,1]$. Finally, the continuity of the map $\lambda \to Q_\lambda$ is clear from the differentiable dependence of geodesics on their endpoints in convex neighborhoods. □

As in Uhlenbeck (1975, p. 79), we denote by $L_* = L \,|\, M_{(p,q)}$ the restriction of the Lorentzian arc length functional to the subset $M_{(p,q)}$ of $C_{(p,q)}$. Just as in the Riemannian case [cf. Milnor (1963, Theorem 16.2)], it can be seen that $L_* : M_{(p,q)} \to \mathbb{R}$ is a faithful model of $L : C_{(p,q)} \to \mathbb{R}$ in the following sense.

Proposition 10.37. *If (M, g) is globally hyperbolic, then the critical points of $L_* = L \,|\, M_{(p,q)}$ are smooth timelike geodesic segments from p to q. These critical points are nondegenerate iff p is not conjugate to q along any timelike geodesic from p to q. Moreover, the index of each critical point is the same in $C_{(p,q)}$ and $M_{(p,q)}$, namely, the index by conjugate points.*

Proof. Identify $M_{(p,q)}$ with k-chains $\{x_1, x_2, \ldots, x_k\}$ from p to q with $x_i \in S_i$ for each i, as in (10.21). Set $x_0 = p$ and $x_{k+1} = q$. Let $\gamma_i : [0,1] \to M$

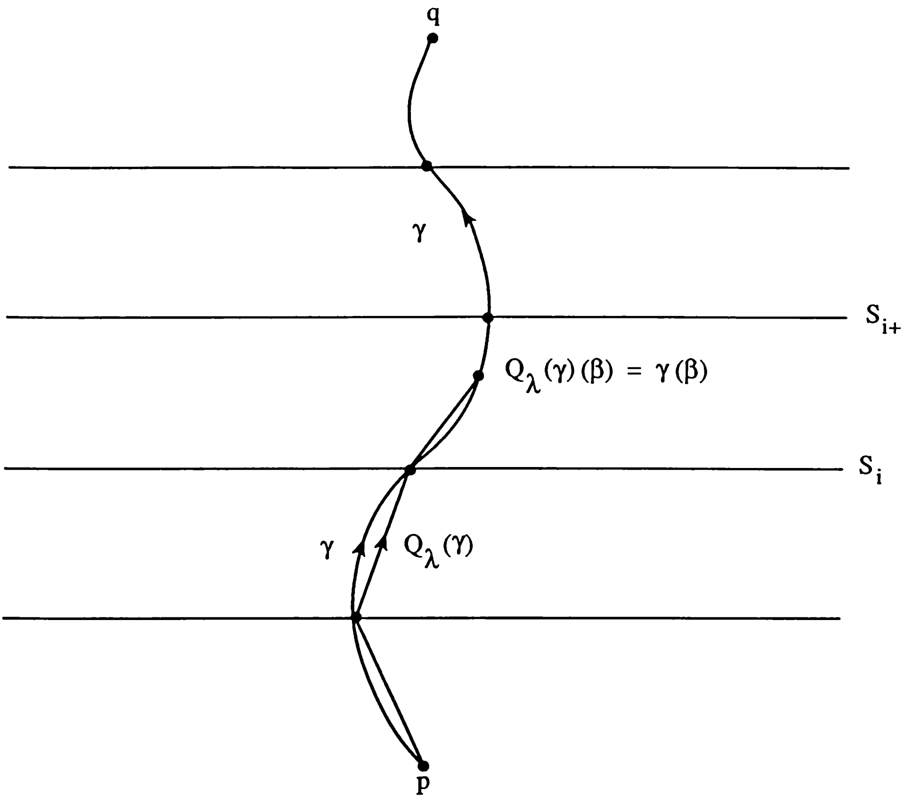


FIGURE 10.2. In the proof of Proposition 10.36, the curve $Q_\lambda(\gamma)$ is used to approximate the given curve γ.

denote the unique maximal timelike geodesic segment from x_i to x_{i+1} for $i = 0, 1, \ldots, k$. Then

$$L_*(x_1, x_2, \ldots, x_k) = \sum_{i=0}^{k+1} \int_0^1 \sqrt{-\langle \gamma_i'(t), \gamma_i'(t) \rangle}\, dt.$$

Suppose we have a smooth deformation $\{x_1(t), x_2(t), \ldots, x_k(t)\}$ of the given chain $\{x_1, x_2, \ldots, x_k\}$. Then since each $t \to x_i(t)$ is a curve in S_i, the variation vector field V of the deformation must lie in $T_{x_i}(S_i)$ at x_i. Also as we are deforming the given chain, which represents a piecewise smooth timelike geodesic, through piecewise smooth timelike geodesics with discontinuities only at the S_i's, the space of deformations of $\{x_1, x_2, \ldots, x_k\}$ may be identified with all vector fields Y along $\{x_1, x_2, \ldots, x_k\}$ so that $Y(x_i) \in T_{x_i}(S_i)$ for $i = 1, 2, \ldots, k$, $Y(p) = Y(q) = 0$, and $Y \mid \gamma_i$ is a smooth Jacobi field along γ_i for each $i = 0, 1, \ldots, k$. By choice of the S_i's as in Lemma 10.34, no $\gamma_i : [0,1] \to M$ has any conjugate points. Thus given any $v_i \in T_{x_i}(S_i)$ and $w_i \in T_{x_{i+1}}(S_{i+1})$, there is a unique Jacobi field J along $\gamma_i : [0,1] \to M$ with $J(0) = v_i$ and $J(1) = w_i$. Thus the space of 1-jets of deformations of the given chain $\{x_1, x_2, \ldots, x_k\}$ may simply be identified with the Cartesian product $T_{x_1}(S_1) \times T_{x_2}(S_2) \times \cdots \times T_{x_k}(S_k)$.

Let $\sigma : [a, b] \to M$ be a smooth future directed timelike curve parametrized so that $\sqrt{-\langle \sigma'(s), \sigma'(s) \rangle} = A$ is constant for all $s \in (a, b)$ and so that $\sigma'(a^+)$ and $\sigma'(b^-)$ are timelike tangent vectors. Let $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ be a smooth variation of σ through nonspacelike curves with variation vector field V. If $\alpha_s : [a, b] \to M$ denotes the curve $\alpha_s(t) = \alpha(t, s)$, then the first variation formula for $L'(0) = (d/ds)L(\alpha_s)|_{s=0}$ is given by

$$L'(0) = -\frac{1}{A}\langle V, \sigma' \rangle \Big|_a^b + \frac{1}{A}\int_a^b \langle V, \nabla_{\sigma'}\sigma' \rangle \Big|_t \, dt. \tag{10.22}$$

Thus if σ is a timelike geodesic, $L'(0) = (-1/A)\langle V, \sigma' \rangle \big|_a^b$.

We now apply (10.22) to calculate the first variation of a proper deformation α of an element $\{x_1, x_2, \ldots, x_k\} \in M_{(p,q)}$. As before, let $\gamma_i : [0,1] \to M$ be timelike geodesics from x_i to x_{i+1} for $i = 0, 1, \ldots, k$. Then $\{x_1, x_2, \ldots, x_k\}$ is represented by the piecewise smooth timelike geodesic $\gamma = \gamma_0 * \gamma_1 * \cdots * \gamma_k$. Let V be the variation vector field of α along γ and set $y_i = V(x_i) \in T_{x_i}(S_i)$. As we mentioned above, the piecewise smooth Jacobi field V along γ may then be identified with $(y_1, y_2, \ldots, y_k) \in T_{x_1}(S_1) \times \cdots \times T_{x_k}(S_k)$. Restricting V to each γ_i and applying formula (10.22), we obtain the first variation formula

$$\delta L_*(y_1, y_2, \ldots, y_k) = \sum_{i=1}^{k} \left(\left\langle y_{i-1}, \frac{\gamma_i'(0)}{A_i} \right\rangle - \left\langle y_i, \frac{\gamma_i'(1)}{B_i} \right\rangle \right) \tag{10.23}$$

where

$$A_i = \sqrt{-\langle {\gamma_i}'(0), {\gamma_i}'(0)\rangle}\,, \qquad B_i = \sqrt{-\langle {\gamma_i}'(1), {\gamma_i}'(1)\rangle} \qquad \text{for each } i,$$

for $L_* = L\,|\,M_{(p,q)}$. From (10.23), it is a standard argument to see that $\delta L_*(y_1, y_2, \ldots, y_k) = 0$ for all $(y_1, y_2, \ldots, y_k) \in T_{x_1}(S_1) \times \cdots \times T_{x_k}(S_k)$ if and only if the tangent vectors

$$\frac{{\gamma_i}'(1)}{\sqrt{-\langle {\gamma_i}'(1), {\gamma_i}'(1)\rangle}} \quad \text{and} \quad \frac{\gamma_{i+1}'(0)}{\sqrt{-\langle \gamma_{i+1}'(0), \gamma_{i+1}'(0)\rangle}}$$

in $T_{x_i}(M)$ have the same projection into the subspace $T_{x_i}(S_i)$ of $T_{x_i}(M)$ for each $i = 1, 2, \ldots, k$. This then implies that $\gamma = \gamma_0 * \gamma_1 * \cdots * \gamma_k$ can be reparametrized to be a smooth timelike geodesic from p to q.

Thus we have seen that the critical points of $L_* : M_{(p,q)} \to \mathbb{R}$ and $L : C_{(p,q)} \to \mathbb{R}$ coincide and are exactly the smooth timelike geodesics from p to q. By our proof of the Timelike Morse Index Theorem (Theorem 10.27) by approximating $V_0^{\perp}(c)$ by spaces of piecewise smooth Jacobi fields (cf. Sublemma 10.26), it follows that the indices of $L_* : M_{(p,q)} \to \mathbb{R}$ and $L : C_{(p,q)} \to \mathbb{R}$ coincide. The Timelike Morse Index Theorem (Theorem 10.27) then implies that a critical point c is degenerate if and only if p and q are conjugate along c (cf. also Proposition 10.13). □

With Proposition 10.37 in hand, the following result may now be established.

Proposition 10.38. *Let (M, g) be globally hyperbolic. If p and q are nonconjugate along any nonspacelike geodesic, then $L : C_{(p,q)} \to \mathbb{R}$ and $L_* : M_{(p,q)} \to \mathbb{R}$ have only a finite number of critical points. In particular, there are only finitely many future directed timelike geodesics from p to q.*

Proof. By Proposition 10.37, it is enough to show that $L_* : M_{(p,q)} \to \mathbb{R}$ has only finitely many critical points. Suppose L_* has infinitely many critical points. Then there would be an infinite sequence of smooth timelike geodesics $\{c_n\}_{n=1}^{\infty}$ from p to q. Since each $\pi_i(M_{(p,q)})$ has compact closure in S_i, the c_n's have a subsequence which must then converge to a timelike or null geodesic c

from p to q. Since c is a limit of an infinite sequence of geodesics from p to q, it follows that p is conjugate to q along c, in contradiction. □

Another interesting property of $L_* : M_{(p,q)} \to \mathbb{R}$ is given by the following result.

Proposition 10.39. *$L_* : M_{(p,q)} \to \mathbb{R}$ assumes its maximum on every component of $M_{(p,q)}$.*

Proof. While $M_{(p,q)}$ is an open submanifold of $S_1 \times S_2 \times \cdots \times S_k$, its closure $\overline{M}_{(p,q)}$ is compact in $S_1 \times S_2 \times \cdots \times S_k$. Also

$$L_*(\{x_1, x_2, \ldots, x_k\}) = \sum_{i=1}^{k} d(x_i, x_{i+1}) \leq d(p, q)$$

where $x_0 = p$ and $x_{k+1} = q$ for any timelike chain $\{x_1, x_2, \ldots, x_k\}$ from p to q. Thus as the Lorentzian distance function is finite valued for globally hyperbolic space–times, it follows that L_* is bounded on each connected component U of $M_{(p,q)}$. Let $\{\gamma_n\}$ be a sequence in the connected component U of $M_{(p,q)}$ with $L_*(\gamma_n) \to \sup\{L_*(\gamma) : \gamma \in U\}$ as $n \to \infty$. By compactness of $\overline{M}_{(p,q)}$ in $S_1 \times S_2 \times \cdots \times S_k$, the sequence $\{\gamma_n\}$ has a limit curve γ_∞ in $\overline{U} \subseteq S_1 \times S_2 \times \cdots \times S_k$. Since $\{\gamma_n\}$ was a maximizing sequence for $L_* \,|\, U$, it follows that $L_*(\gamma_\infty) \geq L_*(\sigma)$ for any $\sigma \in \overline{U}$, and $L_*(\gamma_\infty) > 0$. If γ_∞ is represented by the element $(x_1, x_2, \ldots, x_k) \in S_1 \times \cdots \times S_k$, we have $p \leqslant x_1 \leqslant x_2 \leqslant \cdots \leqslant x_k \leqslant q$, and if γ_i is the unique maximal geodesic from x_i to x_{i+1} guaranteed by Lemma 10.34, then each γ_i is either null or timelike. If some γ_i is null, it follows from the first variation (10.23) that γ_∞ may be deformed to a curve $\gamma \in \overline{U}$ with $L_*(\gamma) > L_*(\gamma_\infty)$. This then contradicts the maximality of $L_* \,|\, \overline{U}$ at γ_∞. Thus, in fact, the limit element $\gamma_\infty \in U$. □

Again, consider $M_{(p,q)}$ as a subset of $S_1 \times \cdots \times S_k$. Let $P_i : T_x(M) \to T_x(S_i)$ denote the orthogonal projection map for any $x \in S_i$ and each $i = 1, 2, \ldots, k$. The set $M_{(p,q)}$ regarded as an open subset of $S_1 \times \cdots \times S_k$ may thus be given a Riemannian metric induced from the Lorentzian metric restricted to the spacelike Cauchy hypersurfaces $S_1, S_2, \ldots, S_k$. It then follows from formula (10.23) that at the point of $M_{(p,q)}$ represented by the broken timelike geodesic

$\gamma = \gamma_0 * \gamma_1 * \cdots * \gamma_k$, with each γ_i parametrized on $[0,1]$ and A_i, B_i as above, the gradient of L_* is given by the formula

$$\operatorname{grad} L_* = \left[P_1\left(\frac{\gamma_1'(0)}{A_1} - \frac{\gamma_0'(1)}{B_0}\right), P_2\left(\frac{\gamma_2'(0)}{A_2} - \frac{\gamma_1'(1)}{B_1}\right), \ldots \right.$$
$$\left. \ldots, P_k\left(\frac{\gamma_k'(0)}{A_k} - \frac{\gamma_{k-1}'(1)}{B_{k-1}}\right)\right].$$

Using this formula, Uhlenbeck (1975, pp. 80–81) then establishes the following lemma.

Lemma 10.40. *Let $\beta : (a,b) \to M_{(p,q)}$ be a maximal integral curve for* $\operatorname{grad} L_*$*. Then $b = \infty$ and $\lim_{t\to\infty} \beta(t)$ lies in the critical set of L_*.*

With the behavior of $\operatorname{grad} L_*$ established, it now follows [Uhlenbeck (1975, p. 81)] that if p and q are nonconjugate along any timelike geodesic, then $L_* : M_{(p,q)} \to \mathbb{R}$ is a Morse function for $M_{(p,q)}$. The main result now follows from this fact and Propositions 10.36 and 10.38.

Theorem 10.41. *Let (M,g) be globally hyperbolic, and let p, q be any pair of points in (M,g) with $p \ll q$ and such that p and q are not conjugate along any nonspacelike geodesic from p to q. Then there are only finitely many future directed timelike geodesics in (M,g) from p to q, and the arc length functional $L : C_{(p,q)} \to \mathbb{R}$ is a homotopic Morse function. Thus if $b > a$ are any two noncritical values of L, then $L^{-1}(-\infty, b)$ is homotopically equivalent to the space $L^{-1}(-\infty, a)$ with a cell attached for each smooth timelike geodesic γ from p to q with $a < L(\gamma) < b$, where the dimension of the attached cell is the (geodesic) index of γ. Thus $C_{(p,q)}$ has the homotopy type of a finite CW-complex with a cell of dimension λ for each smooth future directed timelike geodesic γ from p to q of index λ.*

Note that in Theorem 10.41 the topology of $C_{(p,q)}$ is not related to the given manifold topology. But Uhlenbeck (1975, Theorem 3) has shown for a class of globally hyperbolic space–times satisfying a metric growth condition [cf. Uhlenbeck (1975, p. 72)] that the homotopy of the loop space of M itself may be calculated geometrically as follows. Let (M,g) be a globally hyperbolic space–time satisfying Uhlenbeck's metric growth condition. Then there is a class of smooth timelike curves $\gamma : [0, \infty) \to M$ with the following property. For any such timelike curve γ, there is a residual set of points $p \in M$ such that the loop space of M is homotopic to a cell complex with a cell for each null geodesic from p to γ, where the dimension of the cell corresponds to the conjugate point index of the geodesic.

It will be clear from the proof of Proposition 10.42 below that the finiteness of the homotopy type of the timelike loop space $C_{(p,q)}$ in Theorem 10.41 follows from the assumption that p is not conjugate to q along any null geodesic together with the fact that $L(\gamma) \leq d(p,q) < \infty$ for all $\gamma \in \Omega_{(p,q)}$. For complete Riemannian manifolds (N, g_0) on the other hand, it is known from work of Serre (1951) that if N is not acyclic [i.e., $H_i(N;\mathbb{Z}) \neq 0$ for some $i > 0$], then the loop spaces $\Omega_{(p,q)}$ are infinite CW-complexes for all $p, q \in N$. Thus if p and q are nonconjugate along any geodesic, there are infinitely many geodesics $c_n : [0,1] \to N$ from p to q. Since p and q are nonconjugate along any geodesic and $L_0(\gamma) \geq d_0(p,q) > 0$ for all $\gamma \in \Omega_{(p,q)}$, it may be seen that $L_0(c_n) \to \infty$ as $n \to \infty$.

For completeness, we now give a different proof from Proposition 10.38 of the existence of only finitely many critical points for $L : C_{(p,q)} \to \mathbb{R}$. Instead of using the finite-dimensional approximation of $C_{(p,q)}$ by $M_{(p,q)}$, we work directly with $C_{(p,q)}$ using the existence of nonspacelike limit curves as in Section 3.3.

Proposition 10.42. *Let (M,g) be globally hyperbolic, and suppose that $p, q \in M$ with $p \ll q$ are chosen such that p and q are not conjugate along any future directed nonspacelike geodesic. Then there are only finitely many timelike geodesics from p to q.*

Proof. Suppose that there are infinitely many future directed timelike geodesic segments $c_n : [0,1] \to M$ in $C_{(p,q)}$. Using Corollary 3.32 and the arguments of Section 3.3 we obtain a nonspacelike geodesic $c : [0,1] \to M$ with $c(0) = p$ and $c(1) = q$ which is a limit curve of the sequence $\{c_n\}$ and such that a subsequence of $\{c_n\}$ converges to c in the C^0 topology on curves. Since a subsequence of the pairwise distinct tangent vectors $\{c_n{}'(0)\}$ converges to $c'(0)$, it follows that q is conjugate to p along c, in contradiction. □

Suppose that it is only assumed that p and q are not conjugate along any timelike geodesic in the hypotheses of Proposition 10.42. If there are infinitely many timelike geodesics $\{c_n\}$ in $C_{(p,q)}$, we then have $L(c_n) \to 0$ as $n \to \infty$, and the limit curve c in the proof of Proposition 10.42 is a null geodesic such that q is conjugate to p along c. In particular, c contains a future null cut point to p (cf. Section 9.2). Thus, Theorem 10.41 or the proof of Proposition 10.42 also yields the following result which applies, in particular, to the Friedmann cosmological models with $p = \Lambda = 0$.

Corollary 10.43. *Suppose that (M, g) is globally hyperbolic and that $p, q \in M$ are chosen such that $p \ll q$, p is not conjugate to q along any timelike geodesic, and the future null cut locus of p in M is empty. Then there are only finitely many timelike geodesics from p to q.*

Proof. If there were infinitely many timelike geodesics from p to q, there would be a null geodesic c from p to q such that q is conjugate to p along c. But then p contains a future null cut point along c, in contradiction. □

If $\dim M = 2$, then (M, g) contains no null conjugate points (cf. Lemma 10.45). Thus we also have the following result.

Corollary 10.44. *Suppose that (M, g) is any two-dimensional globally hyperbolic space–time and that $p, q \in M$ are chosen such that $p \ll q$ and q is not conjugate to p along any timelike geodesic. Then there are only finitely many timelike geodesics from p to q.*

10.3 The Null Morse Index Theory

This section is devoted to the proof of a Morse Index Theorem for null geodesic segments $\beta : [a, b] \to (M, g)$ in an arbitrary space–time. The appropriate index form we will use, however, is not the standard index form defined on piecewise smooth vector fields orthogonal to β' but rather its projection to the quotient bundle formed by identifying vector fields differing by a multiple of β'. The idea to use the quotient bundle as the domain of definition for the null index form is implicitly contained in the discussion of variation of arc length for null geodesics in Hawking and Ellis (1973, Section 4.5) and is further developed in Bölts (1977) [cf. Robinson and Trautman (1983)]. The first part of this section develops the basic theory of the index form along the lines of Chapters 2 and 4 of Bölts (1977). In the second part of this section, we give a detailed proof of the Morse Index Theorem for null geodesics sketched in Beem and Ehrlich (1979d).

Instead of working with the energy functional, Uhlenbeck (1975, Theorem 4.5) has constructed a Morse theory for nonspacelike curves in globally hyperbolic space–times as follows. Choosing a globally hyperbolic splitting $M = S \times (a, b)$ as in Theorem 3.17, Uhlenbeck projected nonspacelike curves $\gamma(t) = (c_1(t), c_2(t))$ onto the second factor and showed that the functional $J(\gamma) = \int [c_2{}'(t)]^2 \, dt$ yielded an index theory.

It should be mentioned at the outset that the null index theory, unlike the timelike index theory, is interesting only if $\dim M \geq 3$ for the following reason.

Lemma 10.45. *No null geodesic β in any two-dimensional Lorentzian manifold has any null conjugate points.*

Proof. Let $\beta : (a, b) \to (M, g)$ be an arbitrary null geodesic. Suppose that J is a Jacobi field along β with $J(t_1) = J(t_2) = 0$ for some $t_1 \neq t_2$ in (a, b). Just as in the proof of Lemma 10.9, we have $\langle J, \beta' \rangle'' = -\langle R(J, \beta')\beta', \beta' \rangle = 0$ so that $\langle J(t), \beta'(t) \rangle = 0$ for all $t \in (a, b)$. Since spacelike and null tangent vectors are never orthogonal when $\dim M = 2$, the space of vector fields Y along β perpendicular to β' is spanned by β' itself. Hence $J(t) = f(t)\beta'(t)$ for some smooth function $f : (a, b) \to \mathbb{R}$. The Jacobi equation then becomes $0 = J'' + R(J, \beta')\beta' = f''\beta' + fR(\beta', \beta')\beta' = f''\beta'$ by the skew symmetry of the curvature tensor in the first two slots. Hence $f''(t) = 0$ for all $t \in (a, b)$. Since $J(t_1) = J(t_2) = 0$, it follows that $f = 0$. Thus $J = 0$ as required. □

In view of Lemma 10.45, we will let (M, g) be an arbitrary space–time of dimension $n \geq 3$ throughout this section and let $\beta : [a, b] \to M$ be a fixed null geodesic segment in M. If we let $V^\perp(\beta)$ denote the $\mathbb{R}$-vector space of all piecewise smooth vector fields Y along β with $\langle Y(t), \beta'(t) \rangle = 0$ for all $t \in [a, b]$, then $\beta'(t)$ and $t\beta'(t)$ are both Jacobi fields in $V^\perp(\beta)$. Further, suppose we consider an index form $I : V_0^{\perp}(\beta) \times V_0^{\perp}(\beta) \to \mathbb{R}$ defined by

$$I(X,Y) = -\int [\langle X', Y'\rangle - \langle R(X,\beta')\beta', Y\rangle]\, dt$$

analogous to the index form (10.1) for timelike geodesics in Section 10.1. Then $\mathcal{A} = \{f(t)\beta'(t) \mid f : [a,b] \to \mathbb{R}$ is a smooth function with $f(a) = f(b) = 0\}$ is an infinite-dimensional vector space such that $I(Y,Y) = 0$ for all $Y \in \mathcal{A}$. Thus the extended index of β defined using the index form $I : V_0^{\perp}(\beta) \times V_0^{\perp}(\beta) \to \mathbb{R}$ is always infinite. Also while $I(f\beta', Y) = 0$ for any $Y \in V_0^{\perp}(\beta)$ and $f\beta' \in \mathcal{A}$, the vector field $f\beta'$ is not a Jacobi field unless $f'' = 0$. Thus the relationships we derived for the index form of a timelike geodesic in Section 10.1 linking Jacobi fields, conjugate points, and the definiteness of the index form fail to hold for $I : V_0^{\perp}(\beta) \times V_0^{\perp}(\beta) \to \mathbb{R}$.

The crux of the difficulty is that β' and $t\beta'$ are both Jacobi fields in $V^{\perp}(\beta)$. Thus by ignoring vector fields in $\mathcal{A}$, it is possible to define an index form for null geodesics nicely related not only to the second variation formula for the energy functional but also to conjugate points and Jacobi fields. This may be accomplished by working with the quotient bundle $V^{\perp}(\beta)/[\beta']$ rather than $V^{\perp}(\beta)$ itself. Using this quotient space, an index form $\overline{I}$ may be defined so that β has no conjugate points if and only if $\overline{I}$ is negative definite [cf. Hawking and Ellis (1973, Proposition 4.5.11), Bölts (1977, Satz 4.5.5)]. Also, a Morse Index Theorem for null geodesic segments in arbitrary space–times may be obtained for the index form $\overline{I}$ [cf. Beem and Ehrlich (1979d)].

Since we are interested in studying conjugate points along null geodesics, it is important to note that Lemma 10.9 and Corollaries 10.10 and 10.11 carry over to the null geodesic case with exactly the same proofs.

Lemma 10.46. *Let $\beta : [a,b] \to (M,g)$ be a null geodesic segment, and let Y be any Jacobi field along β. Then $\langle Y(t), \beta'(t)\rangle$ is an affine function of t. Thus if $Y(t_1) = Y(t_2) = 0$ for distinct $t_1, t_2 \in [a,b]$, then $\langle Y(t), \beta'(t)\rangle = 0$ for all $t \in [a,b]$.*

Accordingly, we may restrict our attention to the following spaces of vector fields.

Definition 10.47. Let $V^\perp(\beta)$ denote the $\mathbb{R}$-vector space of all piecewise smooth vector fields Y along β with $\langle Y(t), \beta'(t)\rangle = 0$ for all $t \in [a,b]$. Let $V_0^\perp(\beta) = \{Y \in V^\perp(\beta) : Y(a) = Y(b) = 0\}$. Also set $N(\beta(t)) = \{v \in T_{\beta(t)}M : \langle v, \beta'(t)\rangle = 0\}$, and let

$$N(\beta) = \bigcup_{a \leq t \leq b} N(\beta(t)).$$

For any $Y \in V^\perp(\beta)$, the vector field $Y' \in V^\perp(\beta)$ may be defined using left-hand limits just as in Section 10.1. Since β is a smooth null geodesic, $\beta'(t) \in N(\beta(t))$ for all $t \in [a,b]$. Thus, following Bölts (1977, pp. 39–44), we make the following algebraic construction. Since $N(\beta(t))$ is a vector space and

$$[\beta'(t)] = \{\lambda\beta'(t) : \lambda \in \mathbb{R}\} \tag{10.24}$$

is a vector subspace of $N(\beta(t))$ for each $t \in [a,b]$, we may define the quotient vector space

$$G(\beta(t)) = N(\beta(t))\,/\,[\,\beta'(t)\,] \tag{10.25}$$

and the quotient bundle

$$G(\beta) = N(\beta)\,/\,[\beta'] = \bigcup_{a \leq t \leq b} G(\beta(t)). \tag{10.26}$$

Elements of $G(\beta(t))$ are cosets of the form $v + [\beta'(t)]$, where $v \in N(\beta(t))$ and the vector subspace $[\beta'(t)]$ is the zero element of $G(\beta(t))$ for each $t \in [a,b]$. Also $v + [\beta'(t)] = w + [\beta'(t)]$ in $G(\beta(t))$ if and only if $v = w + \lambda\beta'(t)$ for some $\lambda \in \mathbb{R}$. We may define a natural projection map $\pi : N(\beta(t)) \to G(\beta(t))$ by

$$\pi(v) = v + [\beta'(t)]. \tag{10.27}$$

The projection map π on each fiber induces a projection map $\pi : N(\beta) \to G(\beta)$ given by $\pi(Y) = Y + [\beta']$, i.e., $\pi(Y)(t) = Y(t) + [\beta'(t)] \in G(\beta(t))$ for each $t \in [a,b]$.

This quotient bundle construction may be given a (non-unique) geometric realization as follows. Let $n \in T_{\beta(0)}M$ be a null tangent vector with

$\langle n, \beta'(0)\rangle = -1$. Parallel translate n along β to obtain a null parallel field η along β with $\langle \eta(t), \beta'(t)\rangle = -1$ for all $t \in [a,b]$. Choose spacelike tangent vectors $e_1, e_2, \ldots, e_{n-2} \in T_{\beta(0)}M$ such that $\langle n, e_j\rangle = \langle \beta'(0), e_j\rangle = 0$ for all $j = 1, 2, \ldots, n-2$ and $\langle e_i, e_j\rangle = \delta_{ij}$ for all $j = 1, 2, \ldots, n-2$. Extend by parallel translation to spacelike parallel vector fields $E_1, E_2, \ldots, E_{n-2} \in V^{\perp}(\beta)$, and set

$$V(\beta(t)) = \left\{ \sum_{j=1}^{n-2} \lambda_j E_j(t) : \lambda_j \in \mathbb{R} \text{ for } 1 \le j \le n-2 \right\}. \tag{10.28}$$

Then $V(\beta(t))$ is a vector subspace of $N(\beta(t))$ consisting of spacelike tangent vectors, and we have a direct sum decomposition

$$N(\beta'(t)) = [\beta'(t)] \oplus V(\beta(t)) \tag{10.29}$$

for each $t \in [a,b]$. Set $V(\beta) = \bigcup_{a \le t \le b} V(\beta(t))$, and let $V_0(\beta) = \{Y \in V(\beta) : Y(a) = Y(b) = 0\}$. If $\beta'(t)$ and $\eta(t)$ are given, $V(\beta)$ is independent of the particular choice of $\{e_1, e_2, \ldots, e_{n-2}\}$ in $T_{\beta(0)}M$. However, if $n \in T_{\beta(0)}M$ and hence η are changed, a different direct sum decomposition (10.29) may arise since the given Lorentzian metric g restricted to $N(\beta(t))$ is degenerate. Nonetheless, we may regard $V(\beta(t))$ as being a geometric realization of the quotient bundle $G(\beta)$ via the map $Z \to Z + [\beta']$ from $V(\beta)$ into $G(\beta)$. It is easily checked that this map is an isomorphism since (10.29) is a direct sum decomposition.

We may define the inverse isomorphism

$$\theta : G(\beta(t)) \to V(\beta(t)) \tag{10.30}$$

for each $t \in [a,b]$ as follows. Given $v \in G(\beta(t))$, choose any $x \in N(\beta(t))$ with $\pi(x) = v$. Decomposing x uniquely as $x = \lambda\beta'(t) + v_0$ with $v_0 \in V(\beta(t))$ by (10.29), set $\theta(v) = v_0$. If any other $x_1 \in N(\beta(t))$ with $\pi(x_1) = v$ had been chosen, we would still have $x_1 = \mu\beta'(t) + v_0$ with the same $v_0 \in V(\beta(t))$ for some $\mu \in \mathbb{R}$. Thus θ is well defined.

As a first step toward defining the index form $\overline{I}$ on the quotient bundle $G(\beta)$, we show how the Lorentzian metric, covariant derivative, and curvature tensor of (M, g) may be projected to $G(\beta)$. First, given any $v, w \in G(\beta(t))$, choose $x, y \in N(\beta(t))$ with $\pi(x) = v$ and $\pi(y) = w$. Define $\overline{g}(v, w)$ by

$$\overline{g}(v, w) = g(x, y). \tag{10.31}$$

Suppose we had chosen $x_1, y_1 \in N(\beta(t))$ with $\pi(x_1) = v$ and $\pi(y_1) = w$. Then $x = x_1 + \lambda\beta'(t)$ and $y = y_1 + \mu\beta'(t)$ for some $\lambda, \mu \in \mathbb{R}$, and thus $g(x, y) = g(x_1, y_1) + \lambda g(y_1, \beta'(t)) + \mu g(x_1, \beta'(t)) + \mu\lambda g(\beta'(t), \beta'(t)) = g(x_1, y_1)$. Hence $\overline{g}$ is well defined. It is also easily checked that $\overline{g}(v, w) = g(\theta(v), \theta(w))$ for all $v, w \in G(\beta(t))$. Hence the metric $\overline{g}$ on $G(\beta(t))$ may be identified with the given Lorentzian metric g on $V(\beta(t))$. Since $g \,|\, V(\beta(t)) \times V(\beta(t))$ is positive definite for each $t \in [a, b]$, the induced metric $\overline{g} : G(\beta(t)) \times G(\beta(t)) \to \mathbb{R}$ thus has the following important property.

Remark 10.48. For each $t \in [a, b]$, the metric $\overline{g} : G(\beta(t)) \times G(\beta(t)) \to \mathbb{R}$ is positive definite.

We now extend the covariant derivative operator acting on vector fields along β to a covariant derivative operator for sections of $G(\beta)$. We first introduce the following notation.

Definition 10.49. Let $\mathfrak{X}(\beta)$ denote the piecewise smooth sections of the quotient bundle $G(\beta)$, and let

$$\mathfrak{X}_0(\beta) = \{W \in \mathfrak{X}(\beta) : W(a) = [\beta'(a)] \text{ and } W(b) = [\beta'(b)]\}.$$

Given $V \in \mathfrak{X}(\beta)$, choose $X \in V^\perp(\beta)$ with $\pi(X) = V$. Set

$$V'(t) = \nabla_{\beta'} V(t) = \pi(\nabla_{\beta'} X(t)). \tag{10.32}$$

If $X_1 \in V^\perp(\beta)$ also satisfies $\pi(X_1) = V$, then $X_1 = X + f\beta'$ for some $f : [a, b] \to \mathbb{R}$, and we obtain $X_1{}' = X' + f'\beta'$ since β is a geodesic. Thus $\pi(X_1{}'(t)) = \pi(X'(t))$ for all $t \in [a, b]$. Hence the covariant derivative for $\mathfrak{X}(\beta)$ given by (10.32) is well defined. It may also be checked that this covariant differentiation is compatible with the metric $\overline{g}$ for $G(\beta)$ and satisfies the usual properties of a covariant derivative.

Given $V \in \mathfrak{X}(\beta)$, choose $X \in V^{\perp}(\beta)$ with $\pi(X) = V$. We then have $X(t) = f(t)\beta'(t) + \theta(V)(t)$ with θ as in (10.30). According to (10.32), we may thus calculate $V'(t)$ using $\theta(V)$ as $V'(t) = \pi(\nabla_{\beta'}(\theta(V))(t))$. Now $\theta(V)$ satisfies $\langle \beta'(t), \theta(V)(t)\rangle = \langle \eta(t), \theta(V)(t)\rangle = 0$ for all $t \in [a,b]$. Since β is a geodesic and η is parallel along β, we obtain on differentiating that $\langle \beta'(t), \nabla_{\beta'}\theta(V)(t)\rangle = \langle \eta(t), \nabla_{\beta'}\theta(V)(t)\rangle = 0$ for all $t \in [a,b]$. Thus $\nabla_{\beta'}\theta(V) \in V(\beta)$. Hence we obtain

$$\theta(V'(t)) = (\theta(V))'(t) \qquad \text{for all } t \in [a,b] \tag{10.33}$$

where the differentiation on the left-hand side is in $G(\beta)$ and on the right-hand side is in $V(\beta)$. Thus if we identify $G(\beta)$ and $V(\beta)$ using θ, covariant differentiation in $G(\beta)$ and in $V(\beta)$ are the same.

To project Jacobi fields and the Jacobi differential equation to $G(\beta)$, it is necessary to define a curvature endomorphism

$$\overline{R}(\,\cdot\,, \beta'(t))\beta'(t) : G(\beta(t)) \to G(\beta(t))$$

for each $t \in [a,b]$. This may be done as follows. Given $v \in G(\beta(t))$, choose any $x \in N(\beta(t))$ with $\pi(x) = v$, and set

$$\overline{R}(v, \beta'(t))\beta'(t) = \pi(R(x, \beta'(t))\beta'(t)). \tag{10.34}$$

This definition is easily seen to be independent of the choice of $x \in N(\beta(t))$ with $\pi(x) = v$ since $R(\beta', \beta')\beta' = 0$. If $v = \pi(x)$ and $w = \pi(y)$ with $x, y \in N(\beta(t))$, it follows from (10.31) and (10.34) that

$$\overline{g}\left(\overline{R}(v, \beta'(t))\beta'(t), w\right) = g\left(R(x, \beta'(t))\beta'(t), y\right). \tag{10.35}$$

Finally from the symmetry properties of $g(R(x,y)z,w)$ we obtain

$$\overline{g}\left(\overline{R}(v, \beta'(t))\beta'(t), w\right) = \overline{g}\left(\overline{R}(w, \beta'(t))\beta'(t), v\right) \tag{10.36}$$

for all $v, w \in G(\beta(t))$.

With these preliminary considerations completed, we are now ready to define Jacobi classes in $G(\beta)$ [cf. Bölts (1977, pp. 43–44)].

Definition 10.50. (*Jacobi Class*) A smooth section $V \in \mathfrak{X}(\beta)$ is said to be a *Jacobi class* in $G(\beta)$ if V satisfies the Jacobi differential equation

$$V'' + \overline{R}(V, \beta')\beta' = [\beta'] \tag{10.37}$$

where the covariant differentiation is given by (10.32) and the curvature endomorphism $\overline{R}$ by (10.34).

Given a Jacobi class $W \in \mathfrak{X}(\beta)$ with $W(a) \neq [\beta'(a)]$ and $W(b) \neq [\beta'(b)]$, we will see in the next series of lemmas that there is a two-parameter family $J_{\lambda,\mu}$ of Jacobi fields of the form $J_{\lambda,\mu} = J + \lambda\beta' + \mu t\beta'$ in $V^{\perp}(\beta)$, $\lambda, \mu \in \mathbb{R}$, with $\pi(J_{\lambda,\mu}) = W$. Nonetheless, it should be emphasized that given a Jacobi class $W \in \mathfrak{X}(\beta)$, there may be *no* Jacobi field J in any geometric realization $V(\beta)$ for $G(\beta)$ with $\pi(J) = W$. On the other hand, there will always exist a Jacobi field $J \in V^{\perp}(\beta)$ with $\pi(J) = W$. But it may be necessary for J to have a component in $[\beta']$. The reason for this is made precise in Lemma 10.52. In the next series of lemmas, we will use the geometric realization $V^{\perp}(\beta) = [\beta'] \oplus V(\beta)$ defined in (10.29).

Lemma 10.51. *Let W be a Jacobi class of vector fields in $G(\beta)$. Then there is a smooth Jacobi field $Y \in V^{\perp}(\beta)$ with $\pi(Y) = W$. Conversely, if Y is a Jacobi field in $V^{\perp}(\beta)$, then $\pi(Y)$ is a Jacobi class in $G(\beta)$.*

Proof. The second part of the lemma is clear, for if $Y'' + R(Y, \beta')\beta' = 0$, then $[\beta'] = \pi(0) = \pi(Y'' + R(Y, \beta')\beta') = (\pi(Y))'' + \overline{R}(\pi(Y), \beta')\beta'$.

It remains to establish the first part of the lemma. Given the Jacobi class W, let Y_1 be a smooth vector field in the geometric realization $V(\beta)$ with $\pi(Y_1) = W$. Since $W'' + \overline{R}(W, \beta')\beta' = [\beta']$ in $\mathfrak{X}(\beta)$, there is a smooth function $f : [a, b] \to \mathbb{R}$ such that $Y_1'' + R(Y_1, \beta')\beta' = f\beta'$. Let $h : [a, b] \to \mathbb{R}$ be a smooth function with $h'' = f$ and set $Y = Y_1 - h\beta'$. Then $\pi(Y) = W$, and

$$\begin{aligned} f\beta' &= Y_1'' + R(Y_1, \beta')\beta' \\ &= (Y + h\beta')'' + R(Y + h\beta', \beta')\beta' \\ &= Y'' + h''\beta' + R(Y, \beta')\beta' \\ &= Y'' + f\beta' + R(Y, \beta')\beta'. \end{aligned}$$

Therefore $0 = Y'' + R(Y, \beta')\beta'$ as required. □

A more precise relationship between Jacobi fields in the geometric realization $V(\beta)$ of $G(\beta)$ and Jacobi classes in $\mathfrak{X}(\beta)$ is given by the following lemma.

Lemma 10.52. *Let W be a Jacobi class in $\mathfrak{X}(\beta)$. Then there is a Jacobi field $J \in V(\beta)$ with $\pi(J) = W$ iff the geometric realization $\theta(W)$ of W in $V(\beta)$ satisfies the condition $R(\theta(W), \beta')\beta'|_t \in V(\beta(t))$ for all $t \in [a, b]$.*

Proof. If $J \in V(\beta)$ and $\pi(J) = W$, we have $\theta(W) = J$. But as J is a Jacobi field, we then obtain

$$R(\theta(W), \beta')\beta' = R(J, \beta')\beta' = -J'' \in V(\beta)$$

since $J \in V(\beta)$.

Now suppose that $R(\theta(W), \beta')\beta' \in V(\beta)$ and $J = \theta(W)$. Then $R(J, \beta')\beta' = R(\theta(W), \beta')\beta'$. Since $W'' + \overline{R}(W, \beta')\beta' = [\beta']$ in $G(\beta)$, we know that J must satisfy a differential equation of the form $J'' + R(J, \beta')\beta' = f\beta'$ in $V^{\perp}(\beta)$. But if $R(\theta(V), \beta')\beta' = R(J, \beta')\beta' \in V(\beta)$, the vector field $J'' + R(J, \beta')\beta'$ is in $V(\beta)$. Hence by the decomposition (10.29), $J'' + R(J, \beta')\beta' = 0$. $\square$

For the purpose of studying conjugate points, it is helpful to prove the following refinement of Lemma 10.51 [cf. Bölts (1977, pp. 43–44)].

Lemma 10.53. *Let $W \in \mathfrak{X}(\beta)$ be a Jacobi class with $W(a) = [\beta'(a)]$ and $W(b) = [\beta'(b)]$. Then there is a unique Jacobi field $Z \in V^{\perp}(\beta)$ with $\pi(Z) = W$ and $Z(a) = Z(b) = 0$.*

Proof. From Lemma 10.51, we know that there exists a Jacobi field $Y \in V^{\perp}(\beta)$ with $\pi(Y) = W$. However, we do not know that $Y(a) = Y(b) = 0$. But for any constants $\lambda, \mu \in \mathbb{R}$, the vector field $Y + \lambda\beta' + \mu t\beta'$ is also a Jacobi field in $V^{\perp}(\beta)$ with $\pi(Y + \lambda\beta' + \mu t\beta') = W$. Since $\pi(Y) = W$, $W(a) = [\beta'(a)]$, and $W(b) = [\beta'(b)]$, we know that $Y(a) = c_1\beta'(a)$ and $Y(b) = c_2\beta'(b)$ for some constants $c_1, c_2 \in \mathbb{R}$. Choosing $\lambda = (c_2 a - c_1 b)(b - a)^{-1}$ and $\mu = b^{-1}[(c_1 b - c_2 a)(b - a)^{-1} - c_2]$, it follows easily that $Z = Y + \lambda\beta' + \mu t\beta'$ satisfies $Z(a) = Z(b) = 0$.

For uniqueness, suppose that Z_1 is a second Jacobi field in $V^{\perp}(\beta)$ with $\pi(Z_1) = W$ and $Z_1(a) = Z_1(b) = 0$. Then $X = Z_1 - Z$ is a Jacobi field of the form $X = h\beta'$ with $X(a) = X(b) = 0$. Since $0 = X'' + R(X, \beta')\beta' = h''\beta' +$

$hR(\beta', \beta')\beta' = h''\beta'$, it follows that h is an affine function. As $h(a) = h(b) = 0$, we must have $h = 0$. Therefore $Z_1 = Z$ as required. $\square$

With Lemma 10.53 in hand, it is now possible to make the following definition. Recall also that if J is any Jacobi field along β with $J(t_1) = J(t_2) = 0$ for $t_1 \neq t_2$, then $J \in V^{\perp}(\beta)$ (cf. Lemma 10.46).

Definition 10.54. (*Conjugate Point*) Let $\beta : [a, b] \to (M, g)$ be a null geodesic. For $t_1 \neq t_2$ in $[a, b]$, t_1 and t_2 are said to be *conjugate along* β if there exists a Jacobi class $W \neq [\beta']$ in $\mathfrak{X}(\beta)$ with $W(t_1) = [\beta'(t_1)]$ and $W(t_2) = [\beta'(t_2)]$. Also $t_1 \in (a, b]$ is said to be a *conjugate point of* β if $t = a$ and t_1 are conjugate along β. Let

$$J_t(\beta) = \{\text{Jacobi fields } Y \text{ along } \beta : Y(a) = Y(t) = 0\}$$

and

$$\begin{aligned}\overline{J}_t(\beta) = \{&\text{Jacobi classes } W \in \mathfrak{X}(\beta) : W(a) = [\beta'(a)] \text{ and} \\ &W(t) = [\beta'(t)]\}.\end{aligned}$$

Then $J_t(\beta) \subseteq V^{\perp}(\beta)$ and t_1 and t_2 are conjugate along β in the sense of Definition 10.54 if and only if there exists a nontrivial Jacobi field J along β with $J(t_1) = J(t_2) = 0$. Also we obtain the following important result from the uniqueness in Lemma 10.53.

Corollary 10.55. *The natural projection map* $\pi : J_t(\beta) \to \overline{J}_t(\beta)$ *is an isomorphism for each* $t \in (a, b]$. *Thus* $\overline{J}_t(\beta)$ *is finite-dimensional, and also* $\dim J_t(\beta) = \dim \overline{J}_t(\beta)$ *for all* $t \in (a, b]$.

We are now ready to study the index form of the null geodesic $\beta : [a, b] \to (M, g)$. For the geometric interpretation of the index form, it will be useful to introduce a functional analogous to the arc length functional for timelike geodesics, namely the energy functional. The reason for using energy rather than arc length is simply that the derivative of the function $f(x) = \sqrt{x}$ does not exist at $x = 0$, but $\sqrt{-g(\beta'(t), \beta'(t))} = 0$ for all $t \in [a, b]$ if β is a null geodesic.

Definition 10.56. (*Energy Function*) Let $\gamma : [c,d] \to (M,g)$ be a smooth nonspacelike curve. The smooth mapping $E_\gamma : [c,d] \to \mathbb{R}$ given by

$$E_\gamma(t) = \frac{1}{2}\int_{s=c}^{t} \|\gamma'(s)\|^2 \, ds = -\frac{1}{2}\int_{s=c}^{t} g(\gamma'(s), \gamma'(s)) \, ds$$

is called the *energy function* of γ, and the number $E(\gamma) = E_\gamma(d)$ is called the *energy* of γ.

The energy of a piecewise smooth nonspacelike curve γ is calculated by summing the energies over the intervals on which γ is smooth just as in formula (4.1) for the arc length functional. If we let $L_\gamma(t) = L(\gamma \,|\, [c,t])$, then the Cauchy–Schwarz inequality yields

$$L_\gamma{}^2(d) \leq 2(d-c)E_\gamma(d). \tag{10.38}$$

Also, equality holds in (10.38) if and only if $\|\gamma'(t)\|$ is constant. Thus equality holds for null or timelike geodesics.

Recall that $V^\perp(\beta)$ denotes the space of piecewise smooth vector fields Y along β that are orthogonal to β', and $V_0^\perp(\beta) = \{Y \in V^\perp(\beta) : Y(a) = 0 \text{ and } Y(b) = 0\}$.

Definition 10.57. (*Index Form*) The *index form* $I : V^\perp(\beta) \times V^\perp(\beta) \to \mathbb{R}$ of the null geodesic $\beta : [a,b] \to M$ is given by

$$I(X,Y) = -\int_a^b \left(\langle X', Y' \rangle - \langle R(X,\beta')\beta', Y \rangle\right) dt \tag{10.39}$$

for any $X, Y \in V^\perp(\beta)$.

Formula (10.39) may be integrated by parts just as in the timelike case (cf. Remark 10.5) to give the alternative but equivalent definition of the index form used in Hawking and Ellis (1973, p. 114) and Bölts (1977, p. 110). Explicitly,

$$I(X,Y) = \int_a^b \langle X'' + R(X,\beta')\beta', Y \rangle \, dt + \sum_{i=0}^{k} \langle \Delta_{t_i}(X'), Y \rangle \tag{10.40}$$

where a partition $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ has been chosen such that $X \mid [t_i, t_{i+1}]$ is smooth for $i = 0, 1, 2, \ldots, k-1$ and such that

$$\Delta_{t_0}(X') = X'(a^+),$$
$$\Delta_{t_k}(X') = -X'(b^-),$$

and

$$\Delta_{t_i}(X') = \lim_{t \to t_i^+} X'(t) - \lim_{t \to t_i^-} X'(t)$$

for $i = 1, 2, \ldots, k-1$.

If α is a variation of the null geodesic β such that the neighboring curves are all null, then all derivatives of the energy functional vanish at $s = 0$ since $E(\alpha_s) = 0$ for all s. Thus it is customary to restrict attention to the following class of variations.

Definition 10.58. (*Admissible Variation*) A piecewise smooth variation $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ of a piecewise smooth nonspacelike curve $\beta : [a, b] \to M$ is said to be *admissible* if all the neighboring curves $\alpha_s : [a, b] \to M$ given by $\alpha_s(t) = \alpha(t, s)$ are timelike for each $s \neq 0$ in $(-\epsilon, \epsilon)$.

Now suppose that for $W \in V^\perp(\beta)$, there exists an admissible variation $\alpha : [a, b] \times (-\epsilon, \epsilon) \to M$ with variation vector field $\alpha_* \, \partial/\partial s|_{(t,0)} = W(t)$ for each $t \in [a, b]$. Let $E(\alpha_s)$ be the energy of the neighboring curve $\alpha_s : [a, b] \to M$ for each $s \in (-\epsilon, \epsilon)$. Then $d/ds\,[E(\alpha_s)]\,|_{s=0} = 0$ since β is a geodesic and

$$\left.\frac{d^2}{ds^2} E(\alpha_s)\right|_{s=0} = I(W, W). \tag{10.41}$$

As $E(\beta) = E(\alpha_0) = 0$, we must have $d^2/ds^2\ [E(\alpha_s)]\,|_{s=0} \geq 0$. Thus a necessary condition for $W \in V^\perp(\beta)$ to be the variation vector field of some admissible variation α of β is that $I(W, W) \geq 0$. Note also that if the future null cut point to $\beta(a)$ along β comes after $\beta(b)$, then there are no admissible proper variations of β (cf. Corollary 4.14 and Section 9.2).

It is immediate from (10.39) that if $X \in V_0^\perp(\beta)$ is a vector field of the form $X(t) = f(t)\beta'(t)$ for any piecewise smooth function $f : [a, b] \to \mathbb{R}$ with $f(a) =$

$f(b) = 0$, and $Y \in V_0^\perp(\beta)$ is arbitrary, then $I(X, Y) = 0$. Thus the index form $I : V_0^\perp(\beta) \times V_0^\perp(\beta) \to \mathbb{R}$ is never negative definite and even worse, always has an infinite-dimensional null space. This suggests that the index form should be projected to $G(\beta)$ [cf. Hawking and Ellis (1973, p. 114), Bölts (1977, p. 111)]. Recall that the notation $\mathfrak{X}(\beta)$ was introduced for piecewise smooth sections of $G(\beta)$, and $\mathfrak{X}_0(\beta) = \{W \in \mathfrak{X}(\beta) : W(a) = [\beta'(a)],\ W(b) = [\beta'(b)]\}$.

Definition 10.59. (*Quotient Bundle Index Form*) The *index form* $\overline{I} : \mathfrak{X}(\beta) \times \mathfrak{X}(\beta) \to \mathbb{R}$ is given by

$$\overline{I}(V, W) = -\int_{t=a}^{b} \left[\overline{g}(V', W') - \overline{g}\left(\overline{R}(V, \beta')\beta', W\right)\right] dt \tag{10.42}$$

where $V, W \in \mathfrak{X}(\beta)$ and $\overline{g}$, $\overline{R}$, and the covariant differentiation of sections of $G(\beta)$ are given by (10.31), (10.34), and (10.32), respectively.

Just as for the index form $I : V^\perp(\beta) \times V^\perp(\beta) \to \mathbb{R}$, formula (10.42) may be integrated by parts to give the formula

$$\overline{I}(V, W) = \int_a^b \overline{g}\left(V'' + \overline{R}(V, \beta')\beta', W\right) dt + \sum_{j=0}^{k} \overline{g}\left(W(t_j), \Delta_{t_j}(V')\right) \tag{10.43}$$

where a partition $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ has been chosen such that $V|[t_i, t_{i+1}]$ is smooth for $i = 0, 1, \ldots, k-1$. In particular, if V is a smooth section of $G(\beta)$ we obtain

$$\overline{I}(V, W) = -\overline{g}(V', W)|_a^b + \int_a^b \overline{g}(V'' + \overline{R}(V, \beta')\beta', W)\, dt, \tag{10.44}$$

and if V is a Jacobi class in $\mathfrak{X}(\beta)$, we have

$$\overline{I}(V, W) = -\overline{g}(V', W)|_a^b. \tag{10.45}$$

Also, since vector fields of the form $f(t)\beta'(t)$ are in the null space of the index form $I : V^\perp(\beta) \times V^\perp(\beta) \to \mathbb{R}$, for any $X, Y \in V^\perp(\beta)$ with $\pi(X) = V$ and $\pi(Y) = W$, it follows that

$$\overline{I}(V, W) = I(X, Y) \tag{10.46}$$

where the index on the left-hand side is calculated in $\mathfrak{X}(\beta)$ and on the right-hand side in $V^{\perp}(\beta)$.

We saw in Section 10.1 that for timelike geodesic segments, the null space of the index form consists precisely of the smooth Jacobi fields vanishing at both endpoints. We now establish the analogous result for the index form $\overline{I}$ on the quotient space $\mathfrak{X}_0(\beta)$ for an arbitrary null geodesic segment $\beta : [a, b] \to M$.

Theorem 10.60. *For piecewise smooth vector classes $W \in \mathfrak{X}_0(\beta)$, the following are equivalent:*

(1) *W is a smooth Jacobi class in $\mathfrak{X}_0(\beta)$.*

(2) *$\overline{I}(W, Z) = 0$ for all $Z \in \mathfrak{X}_0(\beta)$.*

Proof. First, (1) $\Rightarrow$ (2) is clear from formula (10.45). For the purpose of showing (2) $\Rightarrow$ (1), fix the unique piecewise smooth vector field Y in the geometric realization $V(\beta)$ for $G(\beta)$ given by (10.28) with $\pi(Y) = W$ and $Y(a) = Y(b) = 0$. Let $a = t_0 < t_1 < \cdots < t_k = b$ be a finite partition of $[a, b]$ such that $Y \mid [t_j, t_{j+1}]$ is smooth for $j = 0, 1, \ldots, k-1$. Let $\rho : [a, b] \to \mathbb{R}$ be a smooth function with $\rho(t_j) = 0$ for each j with $0 \leq j \leq k$, and $\rho(t) > 0$ otherwise. Set $Z = \pi(\rho(Y'' + R(Y, \beta')\beta')) \in \mathfrak{X}_0(\beta)$. Then from (10.43) we obtain

$$0 = \overline{I}(W, Z) = \int_{t=a}^{b} \rho(t) \cdot \overline{g}(W'' + \overline{R}(W, \beta')\beta', W'' + \overline{R}(W, \beta')\beta')\big|_t \, dt.$$

Remembering that $\overline{g}$ is positive definite (cf. Remark 10.48), we obtain $W''(t) + \overline{R}(W(t), \beta'(t))\beta'(t) = [\beta'(t)]$ except possibly at the t_j's. Thus $W \mid [t_j, t_{j+1}]$ is a smooth Jacobi class for each j. To show that W fits together at the t_j's to form a smooth Jacobi class on all of $[a, b]$, it is enough to show that the vector field Y in the geometric realization $V(\beta)$ representing W is a C^1 vector field. First observe that since $W''(t) + \overline{R}(W(t), \beta'(t))\beta'(t) = [\beta'(t)]$ except possibly at the t_j's, we have using (10.43) that

$$0 = \overline{I}(W, Z) = \sum_{j=1}^{k-1} \overline{g}\left(Z(t_j), \Delta_{t_j}(W')\right)$$

for any $Z \in \mathfrak{X}_0(\beta)$. Since $Y \in V(\beta)$, we have $\langle Y(t), \beta'(t)\rangle = \langle Y(t), \eta(t)\rangle = 0$ for all $t \in [a, b]$. Thus $\langle Y'(t), \beta'(t)\rangle = \langle Y'(t), \eta(t)\rangle = 0$ for $t \notin \{t_0, t_1, \ldots, t_k\}$, and it follows by continuity that $\Delta_{t_j}(Y') \in V(\beta(t_j))$ for each $j = 1, 2, \ldots, k-1$. Let $X \in V(\beta)$ be a smooth vector field with $X(a) = X(b) = 0$ and $X(t_j) = \Delta_{t_j}(Y')$ for each $j = 1, 2, \ldots, k-1$. Set $Z = \pi(X) \in \mathfrak{X}_0(\beta)$. Then we obtain

$$0 = \overline{I}(W, Z) = \sum_{j=1}^{k-1} \langle \Delta_{t_j}(Y'), \Delta_{t_j}(Y') \rangle$$

which yields $\Delta_{t_j}(Y') = 0$ for $j = 1, 2, \ldots, k-1$. Hence Y' is continuous at the t_j's as required. □

Notice that in the first part of the proof of (2) ⇒ (1) of Theorem 10.60, we do *not* know that $Y'' + R(Y, \beta')\beta' \in V(\beta)$ even though $Y \in V(\beta)$. Thus we cannot conclude that Y is a Jacobi field except at the t_j's. This is precisely the point in passing to the quotient bundle $G(\beta)$ in which $W'' + \overline{R}(W, \beta')\beta'$ lies in $G(\beta)$ by construction and the induced metric $\overline{g}$ is positive definite on $G(\beta) \times G(\beta)$.

The aim of the next portion of this section is to prove the important result that the index form $\overline{I} : \mathfrak{X}_0(\beta) \times \mathfrak{X}_0(\beta) \to \mathbb{R}$ is negative definite if and only if there are no conjugate points to $t = a$ along β in $[a, b]$. A proof of this result, which is implicitly given in Propositions 4.5.11 and 4.5.12 of Hawking and Ellis (1973), is given in complete detail in Bölts (1977, pp. 117–123). Because the proof of negative definiteness differs considerably from the corresponding proof (cf. Theorem 10.22) for the timelike geodesics, we briefly outline the proof in the timelike case in order to clarify the differences. Recall that we first showed that arbitrary proper piecewise smooth variations α of a timelike geodesic segment have timelike neighboring curves α_s for sufficiently small s. It was then a consequence of the Gauss Lemma and the inverse function theorem that if $c : [a, b] \to M$ is a future directed timelike geodesic segment without conjugate points, then for any proper piecewise smooth variation α of c, the neighboring curves α_s satisfy $L(\alpha_s) \leq L(c)$ for all s sufficiently small. This implied that the index form is negative semidefinite in the absence of conjugate points. Algebraically, we were then able to prove the negative definiteness. Conversely, if c had a conjugate point in $(a, b]$, we produced a piecewise smooth vector field $Z \in V_0^{\perp}(c)$ of zero index using a nontrivial Jacobi field guaranteed by the existence of a conjugate point to $t = a$.

For null geodesic segments, on the other hand, it is necessary to work with admissible variations to get a sensible theory. But from the theory of null cut points we know that if $\beta(b)$ comes before the future null cut point of $\beta(a)$ along β, then there are no future directed timelike curves in M from $\beta(a)$ to $\beta(b)$. Thus the geometric argument of lifting in the absence of conjugate points and using the Gauss Lemma cannot be used to obtain the negative semidefiniteness of the index form $\overline{I}$ in the absence of conjugate points. Instead, it is necessary to work directly with the Jacobi fields themselves. This is most conveniently done by using Jacobi tensors.

By contrast, the proof of the result that if $\overline{I} < 0$, then there are no conjugate points may be done just as in the Riemannian and timelike geodesic cases. Considerably more complicated proofs are presented in Hawking and Ellis (1973, Proposition 4.5.12) and Bölts (1977, Satz 4.5.3) because these authors wish to obtain the result that if $\beta : [a, b] \to M$ has a conjugate point $\beta(t)$ to $\beta(a)$ with $t \in (a, b)$, then there is a timelike curve from $\beta(a)$ to $\beta(b)$ "close" to β. The variation vector field W of the proper admissible variation of β used to prove this result projects to a vector field $\overline{W} = \pi(W) \in \mathfrak{X}_0(\beta)$, $\overline{W} \neq [\beta']$, which satisfies $\overline{I}(\overline{W}, \overline{W}) \geq 0$ by (10.46) and hence demonstrates the remaining half of Theorem 10.69. For completeness, we will also give the proof of this Proposition 4.5.12 of Hawking and Ellis. Notice that this result proves that the null cut point of β comes at or before the first null conjugate point of β.

In order to prove that $\overline{I}$ is negative definite if β has no conjugate points, we now provide a brief description of the theory of Jacobi and Lagrange tensors [cf. Bölts (1977, pp. 45–49)]. A description of these tensors from the point of view of the ∇ notation of this monograph has been given in Eschenburg (1975) for Riemannian manifolds and also in Eschenburg and O'Sullivan (1976) where these tensors were used to study the divergence of geodesics in complete Riemannian manifolds [cf. also Green (1958)]. Recall that given the null geodesic $\beta : [a, b] \to M$, we chose a null parallel vector field η along β with $g(\eta(t), \beta'(t)) = -1$ for all $t \in [a, b]$ and then chose spacelike parallel fields $\{E_1, \ldots, E_{n-2}\}$ along β satisfying $g(E_i(t), E_j(t)) = \delta_{ij}$ and $g(E_i(t), \beta'(t)) = g(E_i(t), \eta(t)) = 0$ for all i, j and t. Let $E_0 = \beta'$ below.

A $(1,1)$ tensor field $A(t)$ of $V^{\perp}(\beta)$ is a linear map

$$A = A(t) : N(\beta(t)) \to N(\beta(t))$$

for each $t \in [a, b]$. Thus if $v \in N(\beta(t))$, then $A(t)(v) \in N(\beta(t))$. A composite endomorphism $RA(t) : N(\beta(t)) \to N(\beta(t))$ may be defined by setting

$$RA(t)(v) = R\big(A(t)(v), \beta'(t)\big)\,\beta'(t) \in N(\beta(t)) \tag{10.47}$$

for any $v \in N(\beta(t))$. The tensor field $A(t)$ may be defined to be smooth (respectively, piecewise smooth) if the maps

$$[a, b] \to V^{\perp}(\beta) \quad \text{given by} \quad t \mapsto A(E_j)(t)$$

for each $j = 1, 2, \ldots, n-2$ are smooth (respectively, piecewise smooth). Writing

$$A(E_j) = \sum_{i=1}^{n-2} f_j^i E_i + f_j \beta' \tag{10.48}$$

where $f_j^i, f_j : [a, b] \to \mathbb{R}$, we may define the $(1,1)$ tensor field $A'(t) : N(\beta(t)) \to N(\beta(t))$ by

$$A'(E_j) = \sum_{i=1}^{n-2} \left(f_j^i\right)' E_i + {f_j}'\beta'$$

for $j = 1, 2, \ldots, n-2$. With these rules, it follows that composition of $(1,1)$ tensors may be identified with matrix multiplication of the component functions f_j^i and f_j. Thus if $A = A(t)$ and $B = B(t)$ are $(1,1)$ tensor fields along β, we have

$$(AB)' = A'B + AB'. \tag{10.49}$$

Applying (10.49) to the formula $AA^{-1} = \mathrm{Id}$ when $A(t) : N(\beta(t)) \to N(\beta(t))$ is nonsingular for all $t \in [a, b]$, we obtain for nonsingular $(1,1)$ tensor fields the differentiation rule

$$(A^{-1})' = -A^{-1}A'A^{-1}. \tag{10.50}$$

For the purpose of the null index theory, however, we need to consider $(1,1)$ tensor fields not on $V^{\perp}(\beta)$ but rather on the quotient bundle $G(\beta)$. A $(1,1)$ tensor field $\overline{A} = \overline{A}(t) : G(\beta(t)) \to G(\beta(t))$ is a linear map for each $t \in [a,b]$ which maps vector classes to vector classes. Using the projection of the covariant derivative to $\mathfrak{X}(\beta)$ [cf. equation (10.32)], we may differentiate piecewise smooth $(1,1)$ tensor fields in $G(\beta(t))$ and obtain the formulas

$$(\overline{A}\,\overline{B})' = \overline{A}'\,\overline{B} + \overline{A}\,\overline{B}' \tag{10.51}$$

and

$$(\overline{A}^{-1})' = -(\overline{A})^{-1}\overline{A}'(\overline{A})^{-1} \tag{10.52}$$

provided $\overline{A}$ is nonsingular. Since the projected metric $\overline{g} : G(\beta(t)) \times G(\beta(t)) \to \mathbb{R}$ is positive definite, we may also define the *adjoint* $\overline{A}^* = \overline{A}^*(t)$ to the $(1,1)$ tensor field $A(t)$ for $G(\beta(t))$ by the formula

$$\overline{g}(\overline{A}(W), Z) = \overline{g}(\overline{A}^*(Z), W) \tag{10.53}$$

for all vector classes $Z, W \in G(\beta(t))$ and all $t \in [a,b]$. We may also define the composed endomorphism $\overline{R}\,\overline{A} : G(\beta(t)) \to G(\beta(t))$ by

$$\overline{R}\,\overline{A}(W) = \overline{R}(\overline{A}(W), \beta')\beta' \tag{10.54}$$

where $\overline{R}$ is the projected curvature operator given by (10.34). Also the *kernel* $\ker\big(\overline{A}(t)\big)$ of the $(1,1)$ tensor field $\overline{A}(t) : G(\beta(t)) \to G(\beta(t))$ is the vector space

$$\ker(\overline{A}(t)) = \{w \in G(\beta(t)) : \overline{A}(t)(w) = [\beta'(t)]\}. \tag{10.55}$$

A $(1,1)$ tensor field $\overline{A}(t) : G(\beta(t)) \to G(\beta(t))$ is the zero tensor field, written $\overline{A} = 0$, if $\overline{A}(t)(w) = [\beta'(t)]$ for all $w \in G(\beta(t))$ and all $t \in [a,b]$.

With these preliminaries out of the way, we are ready to define Jacobi and Lagrange tensor fields.

Definition 10.61. (*Jacobi Tensor Field*) A smooth $(1,1)$ tensor field $\overline{A}: G(\beta) \to G(\beta)$ is said to be a *Jacobi tensor* if $\overline{A}$ satisfies the conditions

$$\overline{A}'' + \overline{R}\,\overline{A} = 0 \tag{10.56}$$

and

$$\ker(\overline{A}(t)) \cap \ker(\overline{A}'(t)) = \{[\beta'(t)]\} \tag{10.57}$$

for all $t \in [a, b]$.

Condition (10.56) has the consequence that if $Y \in \mathfrak{X}(\beta)$ is any parallel vector class along β, then $J = \overline{A}(Y)$ is a Jacobi class along β. This follows since

$$\begin{aligned} J'' + \overline{R}(J, \beta')\beta' &= \overline{A}''(Y) + \overline{R}\,\overline{A}(Y) \\ &= (\overline{A}'' + \overline{R}\,\overline{A})(Y) = 0 \end{aligned}$$

using $Y' = 0$. Condition (10.57) has the following implication: if $Y_1, \ldots, Y_{n-2}$ are linearly independent parallel sections of $\mathfrak{X}(\beta)$, then $\overline{A}(Y_1), \ldots, \overline{A}(Y_{n-2})$ are linearly independent Jacobi sections in the following sense. If $\lambda_1, \ldots, \lambda_{n-2}$ are real constants such that

$$\sum_{j=1}^{n-2} \lambda_j \overline{A}(Y_j)(t) = [\beta'(t)]$$

for all $t \in [a, b]$, then $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-2} = 0$.

The converse of the above construction may be used to show that Jacobi tensors exist. For let $E_1, E_2, \ldots, E_{n-2}$ be the spacelike parallel fields spanning $V(\beta)$ chosen in (10.28). Let $\overline{E}_j = \pi(E_j)$ be the corresponding parallel vector class in $\mathfrak{X}(\beta)$. Also let $\overline{J}_1, \overline{J}_2, \ldots, \overline{J}_{n-2}$ be the Jacobi classes along β with initial conditions $\overline{J}_i(a) = [\beta'(a)]$ and $\overline{J}_i{}'(a) = \overline{E}_i(a)$ for $i = 1, 2, \ldots, n-2$. A Jacobi tensor $\overline{A}$ satisfying the initial conditions $\overline{A}(a) = 0$, $\overline{A}'(a) = \mathrm{Id}$ may then be defined by requiring that

$$\overline{J}_i = \overline{A}(\overline{E}_i) \tag{10.58}$$

for each $i = 1, 2, \ldots, n-2$ and extending to all $G(\beta)$ by linearity. Since the $\overline{J}_i$'s are Jacobi classes and the $\overline{E}_i$'s are parallel classes in $\mathfrak{X}(\beta)$, it follows that $\overline{A}$ satisfies (10.56). To check that $\overline{A}$ satisfies (10.57), suppose that $\overline{A}(t)(\overline{v}) = \overline{A}'(t)(\overline{v}) = [\beta'(t)]$ for some $\overline{v} \in G(\beta(t))$. Choose the unique $v \in V(\beta(t))$ with $\pi(v) = \overline{v}$, and write $v = \sum_{i=1}^{n-2} \lambda_i E_i(t)$. Then $\overline{v} = \sum_{i=1}^{n-2} \lambda_i \overline{E}_i(t)$, and we obtain $[\beta'(t)] = \overline{A}(\overline{v}) = \sum_{i=1}^{n-2} \lambda_i \overline{J}_i(t)$ and

$$[\beta'(t)] = \overline{A}'(\overline{v}) = \sum_{i=1}^{n-2} \lambda_i \overline{A}'(\overline{E}_i)\Big|_t = \sum \lambda_i \overline{J}_i{}'(t).$$

Thus $\overline{J} = \sum_{i=1}^{n-2} \lambda_i \overline{J}_i$ is a Jacobi class in $G(\beta)$ with $\overline{J}(t) = \overline{J}'(t) = [\beta'(t)]$. Hence $\overline{J} = [\beta']$. Thus $\sum_{i=1}^{n-2} \lambda_i \overline{E}_i(s) = [\beta'(s)]$ for all $s \in [a, b]$, contradicting the linear independence of the parallel classes $\overline{E}_1, \overline{E}_2, \ldots, \overline{E}_{n-2}$. Therefore $\ker(\overline{A}(s)) \cap \ker(\overline{A}'(s)) = [\beta'(s)]$ for all $s \in [a, b]$ as required.

It is also useful to note the following result.

Lemma 10.62. $\ker(\overline{A}(s_0)) \cap \ker(\overline{A}'(s_0)) = [\beta'(s_0)]$ *for some* $s_0 \in [a, b]$ *iff* $\overline{A}$ *satisfies condition* (10.57).

Proof. Suppose $v \in \ker(\overline{A}(s_0)) \cap \ker(\overline{A}'(s_0))$, $v \neq [\beta'(s_0)]$. Let Y be the parallel class in $\mathfrak{X}(\beta)$ with $Y(s_0) = v$. Then $J = \overline{A}(Y)$ is a Jacobi class satisfying $J(s_0) = \overline{A}(v) = [\beta'(s_0)]$ and $J'(s_0) = \overline{A}'(v) = [\beta'(s_0)]$. Hence $J = [\beta']$ from which it follows that $Y(t) \in \ker(\overline{A}(t)) \cap \ker(\overline{A}'(t))$ for all $t \in [a, b]$. □

The following two lemmas are not difficult to obtain from the relationship between parallel classes, Jacobi classes, and Jacobi tensors indicated above [cf. Bölts (1977), p. 28 and p. 49)].

Lemma 10.63. *The points* $\beta(t_0)$ *and* $\beta(t_1)$ *are conjugate along the null geodesic* $\beta : [a, b] \to M$ *iff there exists a Jacobi tensor* $\overline{A} : G(\beta) \to G(\beta)$ *with* $\overline{A}(t_0) = 0$, $\overline{A}'(t_0) = Id$, *and* $\ker(\overline{A}(t_1)) \neq \{[\beta'(t_1)]\}$.

Lemma 10.64. *Let* $\beta : [a, b] \to M$ *be a null geodesic without conjugate points to* $t = a$. *Then there exists a unique smooth* $(1, 1)$ *tensor field* $\overline{A} : G(\beta) \to G(\beta)$ *satisfying the differential equation* $\overline{A}'' + \overline{R}\,\overline{A} = 0$ *with* *given boundary conditions* $\overline{A}(a) : G(\beta(a)) \to G(\beta(a))$ *and* $\overline{A}(b) : G(\beta(b)) \to G(\beta(b))$.

Proof. Let $\mathcal{J}$ denote the space of $(1,1)$ tensor fields in $G(\beta)$ satisfying the differential equation $\overline{A}'' + \overline{R}\,\overline{A} = 0$. A linear map $\phi : \mathcal{J} \to L(G(\beta(a))) \times L(G(\beta(b)))$ may then be defined by $\phi(\overline{A}) = (\overline{A}(a), \overline{A}(b))$. Since $\beta : [a,b] \to M$ has no conjugate points, ϕ is injective. For if $\phi(\overline{A}) = (0,0)$ and Y is any parallel vector class in $\mathfrak{X}(\beta)$, then $J = \overline{A}(Y)$ is a Jacobi class with $J(a) = [\beta'(a)]$ and $J(b) = [\beta'(b)]$ so that $J = [\beta']$. Since ϕ is an injective linear map and $\dim \mathcal{J} = \dim[\,L(G(\beta(a))) \times L(G(\beta(b)))\,]$, it follows that ϕ is surjective. □

Even though $\overline{R} = \overline{R}^*$ and $(\overline{A}')^* = (\overline{A}^*)'$, the adjoint $\overline{A}^*$ of a Jacobi tensor $\overline{A}$ is not necessarily a Jacobi tensor since $(\overline{R}\,\overline{A})^* \neq \overline{R}\,\overline{A}^*$ in general. Rather, $(\overline{R}\,\overline{A})^* = \overline{A}^*\overline{R}$. Nonetheless, Jacobi tensors and their adjoints have the following useful property which is conveniently stated using the Wronskian tensor field [cf. Eschenburg and O'Sullivan (1976, p. 226), Bölts (1977, p. 23)].

Definition 10.65. (*Wronskian*) Let $\overline{A}$ and $\overline{B}$ be two Jacobi tensors along $G(\beta)$. Then their *Wronskian* $W(\overline{A}, \overline{B})$ is the $(1,1)$ tensor field along $G(\beta)$ given by

$$W(\overline{A}, \overline{B}) = (\overline{A}')^*\overline{B} - \overline{A}^*\overline{B}'. \tag{10.59}$$

It follows from the fact that $\overline{R}^* = \overline{R}$ and equations (10.51) and (10.56) that if $\overline{A}$ and $\overline{B}$ are any two Jacobi tensor fields along $G(\beta)$, then $[W(\overline{A}, \overline{B})]' = 0$. Thus $W(\overline{A}, \overline{B})$ is a constant tensor field. It is then natural to consider the following subclass of Jacobi tensors.

Definition 10.66. (*Lagrange Tensor Field*) A Jacobi tensor field $\overline{A}$ along $G(\beta)$ is said to be a *Lagrange tensor field* if $W(\overline{A}, \overline{A}) = 0$.

For the proof of Proposition 10.68, we need the following consequence of Definition 10.66.

Lemma 10.67. *Let* $\overline{A}$ *be a Jacobi tensor along* $G(\beta)$. *If* $\overline{A}(s_0) = 0$ *for some* $s_0 \in [a,b]$, *then* $\overline{A}$ *is a Lagrange tensor, and in particular,*

$$(\overline{A}')^*\overline{A} = \overline{A}^*\overline{A}'. \tag{10.60}$$

Proof. We know that $W(\overline{A}, \overline{A})$ is a constant tensor already. But if $\overline{A}(s_0) = 0$, then $\overline{A}^*(s_0) = 0$ also, and hence $W(\overline{A}, \overline{A})(s_0) = 0$. Thus $W(\overline{A}, \overline{A}) = 0$ as required. □

We are now ready to prove the following important proposition.

Proposition 10.68. *Let $\beta : [a, b] \to M$ be a null geodesic without conjugate points to $t = a$ in $(a, b]$. Then $\overline{I}(W, W) < 0$ for all $W \in \mathfrak{X}_0(\beta)$, $W \neq [\beta']$.*

Proof. Let $\overline{A}$ be the Jacobi tensor along $G(\beta)$ with initial conditions $\overline{A}(a) = 0$ and $\overline{A}'(a) = \mathrm{Id}$. Since β has no conjugate points, Lemma 10.63 implies that $\ker(\overline{A}(t)) = \{[\beta'(t)]\}$ for all $t \in (a, b]$.

Now let $W \in \mathfrak{X}_0(\beta)$ be arbitrary. Since $\overline{A}$ is nonsingular in $(a, b]$ and $W(a) = [\beta'(a)]$, we may find $Z \in \mathfrak{X}(\beta)$ with $W = \overline{A}(Z)$. By the rules for covariant differentiation of tensor fields, we obtain

$$\overline{A}'(Z) = [\overline{A}(Z)]' - \overline{A}(Z') \tag{10.61}$$

and

$$\overline{A}''(Z) = (\overline{A}'(Z))' - \overline{A}'(Z'). \tag{10.62}$$

Now we are ready to calculate $\overline{I}(W, W)$. First

$$\begin{aligned} I(W, W) &= -\int_a^b [\overline{g}(W', W') - \overline{g}(\overline{R}(W, \beta')\beta', W)]\, dt \\ &= -\int_a^b \left[\overline{g}\left([\overline{A}(Z)]', [\overline{A}(Z)]'\right) - \overline{g}\left(\overline{RA}(Z), \overline{A}(Z)\right)\right] dt \\ &= -\int_a^b [\overline{g}(\overline{A}'(Z), \overline{A}'(Z)) + 2\overline{g}(\overline{A}'(Z), \overline{A}(Z')) \\ &\qquad + \overline{g}(\overline{A}(Z'), \overline{A}(Z')) + \overline{g}(\overline{A}''(Z), A(Z))]\, dt \end{aligned}$$

using (10.56) and (10.61). Now

$$\begin{aligned} \overline{g}(\overline{A}''(Z), \overline{A}(Z)) &= \overline{g}([\overline{A}'(Z)]', \overline{A}(Z)) - \overline{g}(\overline{A}'(Z'), \overline{A}(Z)) \\ &= [\overline{g}(\overline{A}'(Z), \overline{A}(Z))]' - \overline{g}(\overline{A}'(Z), [\overline{A}(Z)]') - \overline{g}(\overline{A}'(Z'), \overline{A}(Z)) \\ &= [\overline{g}(\overline{A}'(Z), \overline{A}(Z))]' - \overline{g}(\overline{A}'(Z), \overline{A}(Z')) - \overline{g}(\overline{A}'(Z), \overline{A}'(Z)) \\ &\qquad - \overline{g}(\overline{A}'(Z'), \overline{A}(Z)). \end{aligned}$$

Substituting into the above formula for $\overline{I}(W,W)$ then yields

$$\begin{aligned}\overline{I}(W,W) = &- \overline{g}\left(\overline{A}'(Z), \overline{A}(Z)\right)\Big|_a^b \\ &- \int_a^b \Big[\overline{g}\left(\overline{A}(Z'), \overline{A}(Z')\right) + \overline{g}\left(\overline{A}'(Z), \overline{A}(Z')\right) \\ &\qquad - \overline{g}\left(\overline{A}'(Z'), \overline{A}(Z)\right)\Big]\, dt \\ = &-\overline{g}\left(\overline{A}'(Z), W\right)\Big|_a^b - \int_a^b \Big[\overline{g}\left(\overline{A}(Z'), \overline{A}(Z')\right) + \overline{g}\left(\overline{A}^*\overline{A}'(Z), Z'\right) \\ &\qquad - \overline{g}\left(Z', (\overline{A}')^*\overline{A}(Z)\right)\Big]\, dt.\end{aligned}$$

Now the first term vanishes since $W(a) = [\beta'(a)]$ and $W(b) = [\beta'(b)]$. Thus we obtain

$$\begin{aligned}\overline{I}(W,W) &= -\int_a^b \left\{\overline{g}\left(\overline{A}(Z'), \overline{A}(Z')\right) + \overline{g}\left([\overline{A}^*\overline{A}' - (\overline{A}')^*\overline{A}](Z), Z'\right)\right\} dt \\ &= -\int_a^b \overline{g}\left(\overline{A}(Z'), \overline{A}(Z')\right) dt\end{aligned}$$

using (10.60). If $Z'(t) = 0$ for all t, then Z is parallel along β and hence $W = \overline{A}(Z)$ would be a nontrivial Jacobi vector class along β with $W(a) = [\beta'(a)]$ and $W(b) = [\beta'(b)]$. Since β has no conjugate points in $(a,b]$, this is impossible. Hence from the positive definiteness of $\overline{g}$ and the nonsingularity of $\overline{A}(t)$ for $t \in (a,b]$, we obtain $\overline{I}(W,W) < 0$ as required. □

This proposition has the well-known geometric consequence in general relativity that if $\beta : [a,b] \to (M,g)$ is free of conjugate points, then no "small" variation of β gives a timelike curve from p to q [cf. Hawking and Ellis (1973, p. 115)].

We are now ready to prove the following theorem.

Theorem 10.69. *Let $\beta : [a,b] \to M$ be a null geodesic segment. Then the following are equivalent:*

(1) *The segment β has no conjugate points to $t = a$ in $(a,b]$.*

(2) $\overline{I}(W,W) < 0$ *for all* $W \in \mathfrak{X}_0(\beta)$, $W \neq [\beta']$.

Proof. We have already shown (1) $\Rightarrow$ (2) in the proof of Proposition 10.68. To show (2) $\Rightarrow$ (1), we suppose $\beta(a)$ is conjugate to $\beta(t_0)$ with $0 < t_0 \leq b$ and produce a nontrivial vector class $W \in \mathfrak{X}_0(\beta)$ with $\overline{I}(W,W) \geq 0$. If $\beta(a)$ is conjugate to $\beta(t_0)$, we know that there is a nontrivial Jacobi class $Z \in \mathfrak{X}(\beta)$ with

$$Z(a) = [\beta'(a)] \qquad \text{and} \qquad Z(t_0) = [\beta'(t_0)].$$

Set

$$W(t) = \begin{cases} Z(t) & a \leq t \leq t_0, \\ [\beta'(t)] & t_0 \leq t \leq b. \end{cases}$$

Then

$$\begin{aligned} \overline{I}(W,W) &= \overline{I}(W,W)_a^{t_0} + \overline{I}(W,W)_{t_0}^b \\ &= -\overline{g}(Z',Z)|_a^{t_0} + 0 = 0 \end{aligned}$$

using (10.45), $Z(a) = [\beta'(a)]$, and $Z(t_0) = [\beta'(t_0)]$. Thus (2) $\Rightarrow$ (1) holds. □

As in the Riemannian and timelike cases (cf. Theorem 10.23), Theorem 10.69 has the following consequence which is essential to the proof of the Morse Index Theorem for null geodesics.

Theorem 10.70 (Maximality of Jacobi Classes). *Let $\beta : [a,b] \to M$ be a null geodesic segment with no conjugate points to $t = a$, and let $J \in \mathfrak{X}(\beta)$ be any Jacobi class. Then for any vector class $Y \in \mathfrak{X}(\beta)$ with $Y \neq J$,*

(10.63) $$Y(a) = J(a), \qquad \text{and} \qquad Y(b) = J(b)$$

we have

(10.64) $$\overline{I}(J,J) > \overline{I}(Y,Y).$$

Proof. This may be established just as in Theorem 10.23 using (10.45) and Theorem 10.69. □

Since the canonical variation (10.5) of Section 10.1 applied to a vector field $W \in V^{\perp}(\beta)$ is *not* necessarily an admissible variation of β in the sense of Definition 10.58, Theorem 10.69 does *not* imply the existence of a timelike curve from $\beta(a)$ to $\beta(b)$ provided that $t = a$ is conjugate to some $t_0 \in (a, b)$ along β. We thus give a separate proof of this result in Theorem 10.72 [cf. Hawking and Ellis (1973, pp. 115–116), Bölts (1977, pp. 117–121)]. It is first helpful to derive conditions for a proper variation of a null geodesic to be admissible. For convenience, we will assume that $\beta : [0, 1] \to M$ so that the formulas in the proof of Theorem 10.72 will be simpler.

Now let $\alpha : [0, 1] \times (-\epsilon, \epsilon) \to M$ be a piecewise smooth proper variation of β such that each neighboring curve $\alpha_s(t) = \alpha(t, s)$ is future timelike for $s \neq 0$. Let $\partial/\partial t$ and $\partial/\partial s$ denote the coordinate vector fields on $[0, 1] \times (-\epsilon, \epsilon)$, and put $V = \alpha_*(\partial/\partial s)$ and $T = \alpha_*(\partial/\partial t)$. Then $T|_{(t,0)} = \beta'(t)$, and $V|_{(t,0)}$ is called the *variation vector field* of the variation α of β. Since α is a proper variation, $V|_{(0,0)} = V|_{(1,0)} = 0$. Also,

$$\frac{d}{ds}(g(T,T))\bigg|_{(t,0)} = 0 \tag{10.65}$$

since $g(T,T)|_{(t,s)} < 0$ for $s \neq 0$ and $g(T,T)|_{(t,0)} = g(\beta'(t), \beta'(t)) = 0$. Calculating,

$$\begin{aligned}\frac{d}{ds}(g(T,T))\bigg|_{(t,0)} &= 2g(\nabla_{\partial/\partial s}T, T)\,|_{(t,0)} = 2g(\nabla_{\partial/\partial t}V, T)|_{(t,0)} \\ &= 2\,\frac{d}{dt}g(V,T)\bigg|_{(t,0)} - 2g(V, \nabla_{\partial/\partial t}T)|_{(t,0)} \\ &= 2\,\frac{d}{dt}(g(V,T))\bigg|_{(t,0)} - 2g(V|_{(t,0)}, \nabla_{\partial/\partial t}\beta'|_t).\end{aligned}$$

Thus since β is a geodesic, we obtain

$$\frac{d}{ds}(g(T,T))\bigg|_{(t,0)} = 2\,\frac{d}{dt}(g(V,T))\bigg|_{(t,0)}.$$

Now $f(t) = g(V,T)|_{(t,0)}$ is a piecewise smooth function with $f(0) = f(1) = 0$. Hence if $f(t) \neq 0$ for some $t \in [0, 1]$, there exists a $t_0 \in (0, 1)$ with $f'(t_0) > 0$. Then

$$\frac{d}{ds}(g(T,T))\bigg|_{(t_0,0)} = 2f'(t_0) > 0$$

in contradiction to (10.65). Hence we obtain as a first necessary condition for α to be an admissible deformation of β that

$$g(V,T)|_{(t,0)} = 0 \tag{10.66}$$

for all $t \in [0,1]$. Thus $V \in V^{\perp}(\beta)$. Consequently

$$\frac{d}{ds}(g(T,T))\bigg|_{(t,0)} = 2\,\frac{d}{dt}(g(V,T))\bigg|_{(t,0)} = 0 \tag{10.67}$$

for all $t \in [0,1]$. It then follows from (10.67) that the neighboring curves α_s of the variation will be timelike provided that the variation vector field satisfies (10.66), (10.67), and the condition that $d^2/ds^2(g(T,T))|_{(t,0)} < -c < 0$ for all $t \in (0,1)$. As above,

$$\frac{d}{ds}(g(T,T)) = 2\frac{d}{dt}(g(V,T)) - 2g(V, \nabla_{\partial/\partial t}T).$$

Hence

$$\begin{aligned}\frac{d^2}{ds^2}(g(T,T))\bigg|_{(t,0)} &= 2\frac{d}{ds}\left[\frac{d}{dt}(g(V,T))\right]\bigg|_{(t,0)} \\ &\quad - 2\,\frac{d}{ds}(g(V,\nabla_{\partial/\partial t}T))\bigg|_{(t,0)} \\ &= 2\frac{d}{dt}\left[\frac{d}{ds}(g(V,T))\right]\bigg|_{(t,0)} - 2\,g(\nabla_{\partial/\partial s}V, \nabla_{\partial/\partial t}T)|_{(t,0)} \\ &\quad -2g(V, \nabla_{\partial/\partial s}\nabla_{\partial/\partial t}T)|_{(t,0)}\,.\end{aligned}$$

Using the identities

$$\begin{aligned}R(V,T)T &= \nabla_{\partial/\partial s}\nabla_{\partial/\partial t}T - \nabla_{\partial/\partial t}\nabla_{\partial/\partial s}T \\ &= \nabla_{\partial/\partial s}\nabla_{\partial/\partial t}T - \nabla_{\partial/\partial t}\nabla_{\partial/\partial t}V\end{aligned}$$

and

$$\nabla_{\partial/\partial t}T\big|_{(t,0)} = \nabla_{\partial/\partial t}\beta'\big|_t = 0,$$

we obtain

$$\begin{aligned}\frac{d^2}{ds^2}(g(T,T))\big|_{(t,0)} &= 2\frac{d}{dt}[g(\nabla_{\partial/\partial s}V,T)+g(V,\nabla_{\partial/\partial s}T)]\big|_{(t,0)} \\ &\quad - 2g(V,\nabla_{\partial/\partial t}\nabla_{\partial/\partial t}V + R(V,T)T)\big|_{(t,0)} \\ &= 2\frac{d}{dt}[g(\nabla_{\partial/\partial s}V,T)+g(V,\nabla_{\partial/\partial t}V)]\big|_{(t,0)} \\ &\quad - 2g(V,\nabla_{\partial/\partial t}\nabla_{\partial/\partial t}V + R(V,T)T)\big|_{(t,0)}.\end{aligned}$$

This calculation yields the following lemma.

Lemma 10.71. *A sufficient condition for the piecewise smooth proper variation $\alpha : [0,1]\times(-\epsilon,\epsilon)\to M$ of the null geodesic $\beta : [0,1]\to M$ to be an admissible variation [i.e., the curves $\alpha_s(t)=\alpha(t,s)$ are timelike for $s\neq 0$] for all $s\neq 0$ sufficiently small is that the variation vector field $V(t)=\alpha_*\partial/\partial s|_{(t,0)}$ satisfies the conditions*

$$g(V(t),\beta'(t))=0 \qquad \text{for all } t\in[0,1]; \tag{10.68}$$

$$\frac{d}{ds}(g(T,T))\big|_{(t,0)}=0 \qquad \text{for all } t\in[0,1];\ \text{and} \tag{10.69}$$

$$\frac{d}{dt}\,[g(\nabla_{\partial/\partial s}V,\beta')+g(V,V')]\,\big|_{(t,0)} - g(V,V''+R(V,\beta')\beta')\,\big|_t < -c < 0 \tag{10.70}$$

for all $t\in(0,1)$ at which V is smooth.

We are now ready to prove the desired result [cf. Hawking and Ellis (1973, p. 115)].

Theorem 10.72. *Let $\beta : [0,1]\to M$ be a null geodesic. If $\beta(t_0)$ is conjugate to $\beta(0)$ along β for some $t_0\in(0,1)$, then there is a timelike curve from $\beta(0)$ to $\beta(1)$.*

Proof. We will suppose that $t_0>0$ is the first conjugate point of $\beta(0)$ along β. It is enough to show that for some t_2 with $t_0<t_2\leq 1$, there is a future directed timelike curve from $\beta(0)$ to $\beta(t_2)$. For then we have $\beta(0)\ll\beta(t_2)\leqslant$

$\beta(1)$ whence $\beta(0) \ll \beta(1)$. Thus there exists a timelike curve from $\beta(0)$ to $\beta(1)$.

Since $\beta(t_0)$ is conjugate to $\beta(0)$ along β, there exists a smooth nontrivial Jacobi class $W \in \mathfrak{X}(\beta)$ with $W(0) = [\beta'(0)]$ and $W(t_0) = [\beta'(t_0)]$. We may write

$$W(t) = f(t)\hat{W}(t) \tag{10.71}$$

where $\hat{W}$ is a smooth vector class along β with $\overline{g}(\hat{W}, \hat{W}) = 1$ and $f : [0, 1] \to \mathbb{R}$ a smooth function. Since t_0 is the first conjugate point along β, $f(0) = f(t_0) = 0$, and changing $\hat{W}$ to $-\hat{W}$ if necessary, we may assume that $f(t) > 0$ for all $t \in (0, t_0)$. Since W is a nontrivial Jacobi class and $W(t_0) = [\beta'(t_0)]$, we have $W'(t_0) \neq [\beta'(t_0)]$. Thus from the formula $W'(t_0) = f'(t_0)\hat{W}(t_0) + f(t_0)\hat{W}'(t_0) = f'(t_0)\hat{W}(t_0)$, we obtain $f'(t_0) \neq 0$. Hence we may choose $t_1 \in (t_0, 1]$ such that $W(t) \neq [\beta'(t)]$ and $f(t) < 0$ for all $t \in (t_0, t_1]$.

With t_1 as above, the idea of the proof is now to show that there exists a $t_2 \in (t_0, t_1]$ such that there is an admissible proper variation $\alpha : [0, t_2] \times (-\epsilon, \epsilon) \to M$ of $\beta \mid [0, t_2]$. Then the neighboring curves α_s of the variation α will be timelike curves from $\beta(0)$ to $\beta(t_2)$ for $s \neq 0$. This will then imply that $\beta(0) \ll \beta(1)$ as required. To this end, we want to deform $W \mid [0, t_1]$ to a vector class $\overline{Z} \mid [0, t_2]$ with $\overline{g}(\overline{Z}, \overline{Z}'' + \overline{R}(\overline{Z}, \beta')\beta') > 0$ so that if $Z \in V^{\perp}(\beta)$ is an appropriate lift of $\overline{Z} \in \mathfrak{X}(\beta \mid [0, t_2])$ and α is a variation of $\beta \mid [0, t_2]$ with variation vector field Z, then conditions (10.69) and (10.70) of Lemma 10.71 will be satisfied.

Consider a vector class of the form

$$\overline{Z}(t) = [b(e^{at} - 1) + f(t)]\hat{W}(t)$$

with $b = -f(t_1)(e^{at_1} - 1)^{-1} \in \mathbb{R}$ and $a > 0$ in $\mathbb{R}$ chosen such that

$$(a^2 + \text{glb}\{h(t) : t \in [0, t_1]\}) > 0,$$

where

$$h(t) = \overline{g}(\hat{W}'' + \overline{R}(\hat{W}, \beta')\beta', \hat{W})\big|_t.$$

Because W is a Jacobi class, we obtain

$$\begin{aligned}0 &= \overline{g}\left(W'' + \overline{R}(W,\beta')\beta', \hat{W}\right)\\ &= f'' + 2f'\,\overline{g}\,(\hat{W}',\hat{W}) + f\,\overline{g}\,(\hat{W}'',\hat{W}) + f\,\overline{g}\left(\overline{R}(\hat{W},\beta')\beta',\hat{W}\right)\\ &= f'' + 2f'\,\overline{g}\,(\hat{W}',\hat{W}) + fh.\end{aligned}$$

But since $\overline{g}(\hat{W}',\hat{W}) = \frac{1}{2}(g(\hat{W},\hat{W}))' = \frac{1}{2}(1)' = 0$, we obtain the formula $f'' = -fh$. Returning to consideration of the vector class $\overline{Z}$, first note that by choice of the constants a and b, we have $\overline{Z}(0) = [\beta'(0)]$ and $\overline{Z}(t_1) = [\beta'(t_1)]$. In view of formula (10.70), we wish to have $\overline{g}(\overline{Z}, \overline{Z}'' + \overline{R}(\overline{Z},\beta')\beta') > 0$ also. Setting $r(t) = b(e^{at} - 1) + f(t)$, remembering that $\overline{g}(\hat{W}',\hat{W}) = 0$, and differentiating yields $\overline{g}(\overline{Z}, \overline{Z}'' + \overline{R}(\overline{Z},\beta')\beta') = r(r'' + rh) = r[be^{at}(a^2+h) - bh + f'' + fh]$. Since $f'' = -fh$, we obtain

$$\overline{g}\,(\overline{Z}, \overline{Z}'' + \overline{R}(\overline{Z},\beta')\beta')|_t = r(t)\,b\,\{e^{at}[a^2 + h(t)] - h(t)\}.$$

Now $b = -f(t_1)(e^{at_1} - 1) > 0$ as $f(t_1) < 0$, so the expression $b\{e^{at}[a^2 + h(t)] - h(t)\} > 0$ for all $t \in [0,t_1]$. Thus $\overline{g}(\overline{Z}, \overline{Z}'' + \overline{R}(\overline{Z},\beta')\beta')|_t > 0$ provided $r(t) > 0$. Since $f(t) > 0$ for $t \in (0,t_0)$, we have $r(t) > 0$ for $t \in (0,t_0]$. By continuity, there is some $t_2 > t_0$ with $r(t) > 0$ for $t \in [t_0,t_2)$ and $r(t_2) = 0$. If $t_2 \geq t_1$, then in fact $t_2 = t_1$ since $r(t_1) = 0$ by construction, and we let $t_2 = t_1$ below. If $t_2 < t_1$, then the vector class $\overline{Z}\,|\,[0,t_2]$ will satisfy $\overline{Z}(t_2) = [\beta'(t_2)]$ since $r(t_2) = 0$ and also $\overline{g}(\overline{Z}, \overline{Z}'' + \overline{R}(\overline{Z},\beta')\beta')|_t > 0$ for all $t \in (0,t_2)$. We now work with $\beta\,|\,[0,t_2]$.

Let $\widetilde{Z} \in V^{\perp}(\beta\,|\,[0,t_2])$ satisfy $\pi(\widetilde{Z}) = \overline{Z}$. Since $\overline{Z}(0) = [\beta'(0)]$ and $\overline{Z}(t_2) = [\beta'(t_2)]$, we have $\widetilde{Z}(0) = \mu\beta'(0)$ and $\widetilde{Z}(t_2) = \lambda\beta'(t_2)$ for some constants $\mu, \lambda \in \mathbb{R}$. Set $Z = \widetilde{Z} - \mu\beta' + [(\mu - \lambda)/t_2]t\beta'$. Then $Z(0) = Z(t_2) = 0$ and $\pi(Z) = \overline{Z}$. Consequently

$$g\,(Z'' + R(Z,\beta')\beta', Z)\,|_t = \overline{g}\left(\overline{Z}'' + \overline{R}(\overline{Z},\beta')\beta', \overline{Z}\right)\Big|_t > 0 \tag{10.72}$$

for all $t \in (0,t_2)$. Choose a constant $\epsilon > 0$ so that

$$\epsilon < \text{glb}\left\{g\,(Z''(t) + R(Z(t),\beta'(t))\beta'(t), Z(t)) : t \in \left[\tfrac{t_2}{4}, \tfrac{3t_2}{4}\right]\right\} \tag{10.73}$$

which is possible in view of (10.72). Now define a function $\rho : [0, t_2] \to \mathbb{R}$ by

$$\rho(t) = \begin{cases} -\epsilon t & 0 \le t \le \frac{t_2}{4}, \\ \epsilon\left(t - \frac{t_2}{2}\right) & \frac{t_2}{4} \le t \le \frac{3t_2}{4}, \\ \epsilon(t_2 - t) & \frac{3t_2}{4} \le t \le t_2. \end{cases}$$

We now have a given vector field $Z \in V_0^{\perp}(\beta \,|\, [0, t_2])$ and a given function $\rho : [0, t_2] \to \mathbb{R}$. Recall that we had fixed a pseudo-orthonormal frame field $E_1, E_2, \ldots, E_{n-2}$, η, β' for β with $\langle \eta, \beta' \rangle = -1$ in (10.28). We now need to find a proper variation $\alpha : [0, t_2] \times (-\epsilon, \epsilon) \to M$ of $\beta \,|\, [0, t_2]$ satisfying the initial conditions

$$\alpha_* \left. \frac{\partial}{\partial s} \right|_{(t,0)} = Z(t) \tag{10.74}$$

and

$$\nabla_{\partial/\partial s} \alpha_* \left. \frac{\partial}{\partial s} \right|_{(t,0)} = [g(Z, Z')|_t - \rho(t)]\eta(t) \tag{10.75}$$

for all $t \in [0, t_2]$. Thus we wish to specify the first and second derivatives of the curves $s \to \alpha(t, s)$ for each $t \in [0, t_2]$. The existence of such a deformation is guaranteed by the theory of differential equations applied to (10.74) and (10.75) written out in terms of a Fermi coordinate system for the geodesic β defined by the pseudo-orthonormal frame $E_1, E_2, \ldots, E_{n-2}$, η, β'.

Given the proper variation α of $\beta \,|\, [0, t_2]$ satisfying (10.74) and (10.75) and setting $T = \alpha_* \partial/\partial t$ and $V = \alpha_* \partial/\partial s$ as above, it follows that

$$g(\nabla_{\partial/\partial s} V, \beta')|_{(t,0)} + g(V, V')|_{(t,0)} = \rho(t).$$

Hence

$$\frac{d}{dt}\left(g(\nabla_{\partial/\partial s} V, \beta') + g(V, V')\right)\big|_{(t,0)} = \rho'(t) = \begin{cases} -\epsilon & 0 \le t < \frac{t_2}{4}, \\ +\epsilon & \frac{t_2}{4} \le t \le \frac{3t_2}{4}, \\ -\epsilon & \frac{3t_2}{4} < t \le t_2. \end{cases}$$

Thus in view of (10.73), the variation α of $\beta \,|\, [0, t_2]$ satisfies condition (10.70) of Lemma 10.71. Applying Lemma 10.71, we find that this variation produces timelike curves α_s from $\beta(0)$ to $\beta(t_2)$ for small $s \neq 0$ as required. □

Corollary 10.73. *The null cut point of $\beta : [0, a) \to M$ comes at or before the first null conjugate point.*

We are at last ready to turn to the proof of the Morse Index Theorem for null geodesics. The proof parallels that of the Timelike Morse Index Theorem, Theorem 10.27. In view of Theorem 10.69, the index $\mathrm{Ind}(\beta)$ and the extended index $\mathrm{Ind}_0(\beta)$ of β with respect to the index form $\overline{I} : \mathfrak{X}_0(\beta) \times \mathfrak{X}_0(\beta) \to \mathbb{R}$ should be defined as follows.

Definition 10.74. (*Index and Extended Index*) The *index* $\mathrm{Ind}(\beta)$ and *extended index* $\mathrm{Ind}_0(\beta)$ of β with respect to the index form $\overline{I} : \mathfrak{X}_0(\beta) \times \mathfrak{X}_0(\beta) \to \mathbb{R}$ are given by

$$\mathrm{Ind}(\beta) = \mathrm{lub}\,\{\dim A : A \text{ is a subspace of } \mathfrak{X}_0(\beta) \text{ and } \overline{I}\,|\,A \times A \text{ is positive definite}\,\}$$

and

$$\mathrm{Ind}_0(\beta) = \mathrm{lub}\,\{\dim A : A \text{ is a subspace of } \mathfrak{X}_0(\beta) \text{ and } \overline{I}\,|\,A \times A \text{ is positive semidefinite}\,\},$$

respectively.

As a preliminary step toward establishing the null index theorem, we need the following lemma.

Lemma 10.75. *If $\beta\,|\,[s,t]$ is free of conjugate points, then given any $\overline{v} \in G(\beta(s))$ and $\overline{w} \in G(\beta(t))$, there is a unique Jacobi class $Z \in \mathfrak{X}(\beta)$ with $Z(s) = \overline{v}$ and $Z(t) = \overline{w}$.*

Proof. Let $v \in V(\beta(s))$ and $w \in V(\beta(t))$ satisfy $\pi(v) = \overline{v}$ and $\pi(w) = \overline{w}$. By the nonconjugacy hypothesis, there is a unique Jacobi field $J \in V^{\perp}(\beta)$ with $J(s) = v$ and $J(t) = w$. Then $Z = \pi(J)$ is a Jacobi class in $\mathfrak{X}(\beta)$ with $Z(s) = \overline{v}$ and $Z(t) = \overline{w}$.

Suppose now that Z_1 is a second Jacobi class in $\mathfrak{X}(\beta)$ with $Z_1(s) = \overline{v}$ and $Z_1(t) = \overline{w}$. By Lemma 10.51, there is a Jacobi field $J_1 \in V^{\perp}(\beta)$ with $\pi(J_1) = Z_1$. Since $Z_1(s) = \overline{v}$ and $Z_1(t) = \overline{w}$, it follows that $J_1(s) = v + c_1\beta'(s)$

and $J_1(t) = w + c_2\beta'(t)$ for some constants $c_1, c_2 \in \mathbb{R}$. It is then elementary to check that constants $\lambda, \mu \in \mathbb{R}$ may be found so that the Jacobi field $J_2(\tau) = J_1(\tau) + \lambda\beta'(\tau) + \mu\tau\beta'(\tau) \in V^\perp(\beta)$ satisfies $J_2(s) = v$ and $J_2(t) = w$. By the nonconjugacy assumption, we must then have $J_2 = J$. But this implies that $Z = \pi(J) = \pi(J_2) = \pi(J_1) = Z_1$ as required. □

We are now ready to establish the finiteness of the index and extended index.

Proposition 10.76. $\mathrm{Ind}_0(\beta)$ *and* $\mathrm{Ind}(\beta)$ *are both finite and are related by the equality*

$$\mathrm{Ind}_0(\beta) = \mathrm{Ind}(\beta) + \dim \overline{J}_b(\beta).$$

Proof. Given the fixed null geodesic $\beta : [a, b] \to M$, choose a finite partition $a = t_0 < t_1 < \cdots < t_k = b$ so that $\beta \,|\, [t_j, t_{j+1}]$ is free of conjugate points for each j. Let $\overline{J}\{t_j\}$ denote the vector space of continuous vector classes W along β such that $W \,|\, [t_j, t_{j+1}]$ is a smooth Jacobi class for each j. It is immediate from the uniqueness in Lemma 10.75 that $\overline{J}\{t_j\}$ is finite-dimensional. We may then define a finite-dimensional approximation $\phi : \mathfrak{X}_0(\beta) \to \overline{J}\{t_j\}$ in analogy with the Riemannian and timelike Lorentzian index theory. Namely, given $W \in \mathfrak{X}_0(\beta)$, let $\phi(W)$ be the piecewise smooth Jacobi class in $\overline{J}\{t_j\}$ such that for each j, the vector class $\phi(W) \,|\, [t_j, t_{j+1}]$ is the unique Jacobi class along $\beta \,|\, [t_j, t_{j+1}]$ with $\phi(W)(t_j) = W(t_j)$ and $\phi(W)(t_{j+1}) = W(t_{j+1})$. Then $\phi \,|\, \overline{J}\{t_j\} = \mathrm{Id}$, and (10.64) applied to each subinterval $[t_j, t_{j+1}]$ yields the inequality

$$\overline{I}(\phi(W), \phi(W)) > \overline{I}(W, W) \tag{10.76}$$

if $W \in \mathfrak{X}_0(\beta)$ and $W \notin \overline{J}\{t_j\}$. Using (10.76), it may be shown just as in the proof of Sublemma 10.26 in section 10.1 that

(10.77) If A is a subspace of $\mathfrak{X}_0(\beta)$ on which $\overline{I} \,|\, A \times A$ is positive semidefinite, then $\phi\,|_A : A \to \overline{J}\{t_j\}$ is injective.

Now let $\mathrm{Ind}'(\beta)$ and $\mathrm{Ind}_0'(\beta)$ denote the index and extended index, respectively, of the index form $\overline{I} : \overline{J}\{t_j\} \times \overline{J}\{t_j\} \to \mathbb{R}$ restricted to $\overline{J}\{t_j\}$. Suppose that A is a subspace of $\mathfrak{X}_0(\beta)$ on which $\overline{I}\,|\,A \times A$ is positive semidefinite. Using (10.76), it is easily seen that $\overline{I} : \overline{J}\{t_j\} \times \overline{J}\{t_j\} \to \mathbb{R}$ is positive semidefinite on the subspace $\phi(A)$ of $\overline{J}\{t_j\}$. Also by (10.77), $\dim A = \dim \phi(A)$. Hence $\mathrm{Ind}_0'(\beta) \geq \mathrm{Ind}_0(\beta)$. On the other hand, as $\overline{J}\{t_j\} \subseteq \mathfrak{X}_0(\beta)$, we have $\mathrm{Ind}_0'(\beta) \leq \mathrm{Ind}_0(\beta)$. Thus $\mathrm{Ind}_0'(\beta) = \mathrm{Ind}_0(\beta)$. Since $\overline{J}\{t_j\}$ is finite-dimensional, these equalities imply that $\mathrm{Ind}(\beta)$ and $\mathrm{Ind}_0(\beta)$ are finite-dimensional.

It remains to show that $\mathrm{Ind}_0(\beta) = \mathrm{Ind}(\beta) + \dim \overline{J}_b(\beta)$. To this end, choose a second finite partition $a = s_0 < s_1 < \cdots < s_{l-1} < s_l = b$ such that $\{s_1, s_2, \ldots, s_{l-1}\} \cap \{t_1, t_2, \ldots, t_{k-1}\} = \emptyset$ and such that $\beta\,|\,[s_i, s_{i+1}]$ has no conjugate points for each i. Then $\overline{J}\{s_i\} \cap \overline{J}\{t_j\} = \overline{J}_b(\beta)$. Thus if $X \in \overline{J}\{s_i\}$ and $X \notin J_b(\beta)$, it follows from (10.76) that

$$\overline{I}(\phi(X), \phi(X)) > \overline{I}(X, X). \tag{10.78}$$

From the preceding part of the proof of this proposition applied to the partition $\{s_i\}$ and associated finite-dimensional vector space $\overline{J}\{s_i\}$, we may choose a vector subspace B_0' of $\overline{J}\{s_i\}$ so that $\overline{I}\,|\,B_0' \times B_0'$ is positive semidefinite and $\mathrm{Ind}_0(\beta) = \dim B_0' < \infty$. Note that $\overline{J}_b(\beta)$ must be a subspace of B_0'. For otherwise there is a nontrivial vector subspace V of $\overline{J}_b(\beta)$ with $V \cap B_0' = \{[\beta']\}$. Then $B_0'' = B_0' \oplus V$ would be a subspace of $\overline{J}\{s_i\}$ on which $\overline{I}\,|\,B_0'' \times B_0''$ is positive semidefinite, but $\dim B_0'' > \dim B_0'$, in contradiction to our choice of B_0'.

Regarding B_0' as a subspace of $\mathfrak{X}_0(\beta)$, we then know from (10.77) that $\phi|_{B_0'} : B_0' \to \overline{J}\{t_j\}$ is injective. Thus if we set $B_0 = \phi(B_0')$, we have $\dim B_0 = \dim B_0' = \mathrm{Ind}_0(\beta)$. Since $\phi\,|\,\overline{J}_b(\beta) = \mathrm{Id}$, we also have that $\overline{J}_b(\beta)$ is a subspace of B_0. Choose a vector subspace B of B_0 such that $B_0 = B \oplus \overline{J}_b(\beta)$. Using (10.78), it may be established just as in the proof of Proposition 10.25 of Section 10.1 that $\overline{I}\,|\,B \times B$ is positive definite. Hence $\mathrm{Ind}'(\beta) \geq \dim B$.

But $\mathrm{Ind}_0(\beta) = \mathrm{Ind}_0'(\beta) = \dim B_0 = \dim B + \dim \overline{J}_b(\beta)$. Thus the proof will be complete if we show that $\mathrm{Ind}'(\beta) \leq \dim B$. Therefore suppose that B' is a subspace of $\overline{J}\{t_i\}$ with $\overline{I}\,|\,B' \times B'$ positive definite and $\mathrm{Ind}'(\beta) = \dim(B') > \dim B$. As $\overline{I}\,|\,B' \times B'$ is positive definite, $B' \cap \overline{J}_b(\beta) = \{[\beta']\}$. Hence $\overline{I}$ is positive semidefinite on the direct sum $B' \oplus \overline{J}_b(\beta)$ so that $\dim B' + \dim \overline{J}_b(\beta) \leq \mathrm{Ind}_0'(\beta)$. On the other hand, $\dim B' + \dim \overline{J}_b(\beta) > \dim B + \dim \overline{J}_b(\beta) = \mathrm{Ind}_0'(\beta)$, in contradiction. Hence $\mathrm{Ind}'(\beta) \leq \dim B$ as required, completing the proof of the proposition. □

Now that we have obtained Proposition 10.76, it is straightforward to see that the proof of Theorem 10.27 of Section 10.1 may be applied to the index form $\overline{I} : G(\beta) \times G(\beta) \to \mathbb{R}$ and the projected positive definite metric $\overline{g}$ to yield the equalities

$$\mathrm{Ind}(\beta) = \sum_{t \in (a,b)} \dim \overline{J}_t(\beta)$$

and

$$\mathrm{Ind}_0(\beta) = \sum_{t \in (a,b]} \dim \overline{J}_t(\beta).$$

Since $\dim \overline{J}_t(\beta) = \dim J_t(\beta)$ by Corollary 10.55, we have thus established the following Morse Index Theorem for null geodesics.

Theorem 10.77. *Let $\beta : [a, b] \to M$ be a null geodesic in an arbitrary space–time. Let $\overline{I} : \mathfrak{X}_0(\beta) \times \mathfrak{X}_0(\beta) \to \mathbb{R}$ be the index form on piecewise smooth sections of the quotient bundle $G(\beta)$ defined in (10.42). Then β has only finitely many conjugate points, and the index* $\mathrm{Ind}(\beta)$ *and extended index* $\mathrm{Ind}_0(\beta)$ *of $\overline{I} : \mathfrak{X}_0(\beta) \times \mathfrak{X}_0(\beta) \to \mathbb{R}$ are related to the geodesic index of the null geodesic β by the formulas*

$$\mathrm{Ind}(\beta) = \sum_{t \in (a,b)} \dim J_t(\beta)$$

and

$$\mathrm{Ind}_0(\beta) = \sum_{t \in (a,b]} \dim J_t(\beta)$$

where $J_t(\beta)$ denotes the vector space of Jacobi fields Y along β with $Y(a) = Y(t) = 0$.

Extensions of Theorem 10.77 to the focal case of a null geodesic segment perpendicular to spacelike endmanifolds have been given in Ehrlich and Kim (1989a,b). Studies of spacelike conjugate points have recently been made in Helfer (1994a,b).

CHAPTER 11

SOME RESULTS IN GLOBAL LORENTZIAN GEOMETRY

In Chapter 11 we apply the techniques of the preceding chapters to obtain Lorentzian analogues of two remarkable results in global Riemannian geometry. The first, the Bonnet–Myers Diameter Theorem, asserts that if a complete Riemannian manifold N has everywhere positive Ricci curvature bounded away from zero, then N is compact, has finite diameter, and has finite fundamental group. The second result, the Hadamard–Cartan Theorem, states that if a complete Riemannian manifold has everywhere nonpositive sectional curvature, then its universal covering manifold is diffeomorphic to $\mathbb{R}^n$ and thus the higher homotopy groups $\pi_i(N, *) = \{e\}$ for $i \geq 2$. In addition, the universal covering space with the pullback Riemannian metric has the property that any two points may be joined by exactly one geodesic, up to reparametrization.

In Section 11.1 we consider the Lorentzian analogue of the Bonnet–Myers Theorem and in so doing, study the timelike diameter of space–times. The *timelike diameter* $\mathrm{diam}(M, g)$ of a space–time (M, g) is given by

$$\mathrm{diam}(M, g) = \sup\{d(p, q) : p, q \in M\}.$$

Classes of space–times with finite timelike diameter, including the "Wheeler universes," have been studied in general relativity [cf. Tipler (1977c, p. 500)].

If a complete Riemannian manifold has finite diameter, it is compact by the Hopf–Rinow Theorem. Even so, all geodesics have *infinite* length as a result of the metric completeness. But for space–times (M, g), since $L(\gamma) \leq d(p, q)$ for all future directed nonspacelike curves γ from p to q, *every* timelike geodesic must satisfy $L(\gamma) \leq \mathrm{diam}(M, g)$. Thus if a space–time (M, g) has finite timelike diameter, all timelike geodesics have finite length and hence are incomplete. In particular, a space–time (M, g) with finite timelike diameter is timelike geodesically incomplete.

Since we have adopted the signature convention $(-,+,\cdots,+)$ for the Lorentzian metric instead of $(+,-,\cdots,-)$, curvature conditions of positive (respectively, negative) sectional curvature for Riemannian manifolds translate as curvature conditions of negative (respectively, positive) timelike sectional curvature for Lorentzian manifolds.

Using the timelike index theory developed in Section 10.1, we obtain the following Lorentzian analogue of the Bonnet–Myers Theorem for complete Riemannian manifolds. Let (M,g) be a globally hyperbolic space–time with either (1) all nonspacelike Ricci curvatures positive and bounded away from zero, or (2) all timelike sectional curvatures negative and bounded away from zero. Then (M,g) has finite timelike diameter.

In Section 11.2 we give Lorentzian versions of two well-known comparison theorems in Riemannian geometry, the index comparison theorem and the Rauch Comparison Theorem. Using the latter of these two results, we are able to give an easy proof (Corollary 11.12) of the basic fact that in a space–time with everywhere nonnegative timelike sectional curvatures, the differential $\exp_{p_*}$ of the exponential map,

$$\exp_{p_*} : T_v(T_pM) \to T_{\exp_p(v)}M$$

is norm nondecreasing on nonspacelike tangent vectors.

In Section 11.3 we consider two analogues of the Hadamard–Cartan Theorem. The first of these is for future one-connected globally hyperbolic space–times. The space–time (M,g) is said to be *future one-connected* if for any $p,q \in M$ with $p \ll q$, any two smooth, future directed timelike curves with endpoints p and q are homotopic through smooth, future directed timelike curves with endpoints p and q. Using the Morse theory of the timelike path space $C_{(p,q)}$ from Section 10.2, it may be shown that if (M,g) is a future one-connected globally hyperbolic space–time with no nonspacelike conjugate points, then given any $p,q \in M$ with $p \ll q$, there is exactly one future directed timelike geodesic segment (up to parametrization) from p to q (cf. Theorem 11.16).

A Lorentzian manifold is said to be *geodesically connected* if each pair of distinct points is joined by at least one geodesic segment. Global hyperbolicity guarantees the existence of at least one (maximal length) geodesic segment joining any two causally related points. However, it does not yield any information about points which fail to be causally related. Furthermore, geodesic completeness does not imply geodesic connectedness for Lorentzian manifolds. In fact, even compact space–times may fail to be geodesically connected (cf. Example 11.23). Sufficient conditions for geodesic connectedness are given in Proposition 11.22 in terms of disprisonment, pseudoconvexity, and a lack of conjugate points along all geodesics. In Theorem 11.25 we find these conditions actually suffice for our second version of the Hadamard–Cartan Theorem. This result yields sufficient conditions for the exponential map at any fixed point to be a diffeomorphism from its domain in the tangent space onto the manifold M.

11.1 The Timelike Diameter

Motivated by the concept of the diameter of a complete Riemannian manifold, the following analogue has been considered for arbitrary space–times [cf. Beem and Ehrlich (1979c, Section 9)].

Definition 11.1. (*Timelike Diameter*) The *timelike diameter* of the space–time (M, g), denoted by $\operatorname{diam}(M, g)$, is defined to be

$$\operatorname{diam}(M, g) = \sup\{d(p, q) : p, q \in M\}.$$

A similar concept has been used by Tipler (1977a, p. 17) in studying singularity theory in general relativity. Physically, the timelike diameter represents the supremum of possible proper times any particle could possibly experience in the given space–time. A space–time of finite timelike diameter is singular (recall Definition 6.3) in a striking way.

Remark 11.2. If $\operatorname{diam}(M, g) < \infty$, then *all* timelike geodesics have length less than or equal to $\operatorname{diam}(M, g)$ and are thus incomplete.

Proof. Suppose that $c : (a, b) \to M$ is a timelike geodesic which satisfies $L(c) > \operatorname{diam}(M, g)$. We may then find $s, t \in (a, b)$, $s < t$, such that

$$L(c\,|\,[s, t]) > \operatorname{diam}(M, g).$$

But then

$$d(c(s), c(t)) \geq L(c\,|\,[s,t]) > \operatorname{diam}(M, g)$$

which is impossible. □

From a physical point of view, the most interesting space–times of finite timelike diameter are the Wheeler universes [cf. Tipler (1977c, p. 500)]. In particular, the "closed" Friedmann cosmological models are examples of Wheeler universes.

For a complete Riemannian manifold (N, g_0), the diameter is finite if the manifold is compact. In this case, we may always find two points of N whose distance realizes the diameter. On the other hand, for space–times with finite timelike diameter, the diameter is never achieved.

Proposition 11.3. *Let (M, g) be an arbitrary space–time, and suppose that there exist $p, q \in M$ such that $d(p,q) = \operatorname{diam}(M, g)$. Then $d(p,q) = \infty$.*

Proof. Suppose $d(p,q) = \operatorname{diam}(M,g) < \infty$. Let $q' \in I^+(q)$ be arbitrary. Then

$$d(p,q') \geq d(p,q) + d(q,q') > d(p,q) = \operatorname{diam}(M,g)$$

in contradiction. □

Recalling that globally hyperbolic space–times satisfy the finite distance condition (cf. Definition 4.6), we have the following corollary to Proposition 11.3.

Corollary 11.4. *(1) The timelike diameter is never realized by any pair of points in a space–time of finite timelike diameter.*
(2) The timelike diameter is never achieved in a globally hyperbolic space–time.

We now prove the Lorentzian analogue (Theorem 11.9) of Bonnet's Theorem and Myers' Theorem for complete Riemannian manifolds [cf. Cheeger and Ebin (1975, pp. 27–28)]. Similar results have been given by Avez (seminar lecture), Flaherty (unpublished), Uhlenbeck (1975, Theorem 5.4 and Corollary 5.5), and Beem and Ehrlich (1979c, Theorem 9.5). Also, Theorem 11.9 is contained implicitly in stronger results using the Raychaudhuri equation needed for singularity theory in general relativity [cf. Section 12.2 or Hawking and Ellis (1973, Section 4.4)].

Definition 11.5. (*Timelike Two-Plane*) A *timelike two-plane* σ is a two-dimensional subspace of T_pM for some $p \in M$ which is spanned by a spacelike tangent vector and a timelike tangent vector.

Recall that the sectional curvature $K(\sigma)$ of the timelike two-plane σ may be calculated by choosing a basis $\{v, w\}$ for σ consisting of a timelike and a spacelike tangent vector and setting

$$K(\sigma) = \frac{g(R(v,w)w,v)}{g(v,v)g(w,w) - [g(v,w)]^2}.$$

Remark 11.6. Rather than considering all sectional curvatures, it is essential to restrict attention to timelike sectional curvatures for the following reason. If (M, g) is a space–time of dimension $n \geq 3$ and the sectional curvature function of (M, g) is either bounded from above for all nonsingular two-planes or bounded from below for all nonsingular two-planes, then (M, g) has constant sectional curvature [Kulkarni (1979)]. Here a two-plane σ is said to be *nonsingular* if $g(v,v)g(w,w) - [g(v,w)]^2 \neq 0$ for some (and hence for any) basis $\{v, w\}$ for σ. On the other hand, space–times with all *timelike* sectional curvatures $K(\sigma) \leq -k^2$ or all timelike sectional curvatures $K(\sigma) \geq k^2$ exist. But Harris (1982a) has shown that if all timelike sectional curvatures are bounded both from above and below, then (M, g) has constant sectional curvature. Thus no obvious analogue for the Lorentzian sectional curvature exists for pinched Riemannian manifolds [cf. Cheeger and Ebin (1975, p. 118) for Riemannian pinching].

With the signature convention $(-, +, \cdots, +)$ used here for Lorentzian metrics, curvature conditions in Riemannian geometry for complete Riemannian manifolds of positive (respectively, negative) sectional curvature tend to correspond to theorems for globally hyperbolic space–times of negative (respectively, positive) sectional curvature. On the other hand, if we change this signature convention to $(+, -, \cdots, -)$ by setting $(M, \hat{g}) = (M, -g)$, then $K(\hat{g}) = -K(g)$. Thus Riemannian theorems for positive (respectively, negative) sectional curvature correspond to Lorentzian theorems for positive (respectively, negative) sectional curvature for $(M, \hat{g})$ [cf. for instance Flaherty (1975a, pp. 395–396) where the convention $(+, -, \cdots, -)$ is used]. But whether g or $\hat{g}$ is chosen as the Lorentzian metric for M, $\mathrm{Ric}(g) = \mathrm{Ric}(\hat{g})$.

It is convenient to isolate part of the proof of Theorem 11.9 in the following proposition. Recall the notation $V_0^\perp(c)$ from Definition 10.1.

Proposition 11.7. *Let (M, g) be an arbitrary space–time of dimension $n \geq 2$. Suppose that (M, g) satisfies either the curvature condition*

(1) *Every timelike plane σ has sectional curvature $K(\sigma) \leq -k < 0$,*

or the curvature condition

(2) $\mathrm{Ric}(g)(v, v) \geq (n-1)k > 0$ *for all unit timelike tangent vectors $v \in TM$.*

Then if $c : [0, b] \to M$ is any timelike geodesic with $L(c) \geq \pi/\sqrt{k}$, the geodesic segment c has a pair of conjugate points.

Proof. Since curvature condition (1) implies curvature condition (2) by taking the trace, we will prove that condition (2) implies the desired conclusion. For convenience, we will suppose that $c : [0, L] \to M$ is parametrized as a unit speed timelike geodesic with length L. Set $E_n(t) = c'(t)$ for all $t \in [0, L]$, and let $\{E_1, E_2, \ldots, E_{n-1}\}$ be $n-1$ spacelike parallel vector fields along c such that $\{E_1(t), E_2(t), \ldots, E_n(t)\}$ forms a Lorentzian orthonormal basis of $T_{c(t)}M$ for each $t \in [0, L]$. Set $W_i(t) = \sin(\pi t/L)E_i(t)$, so that $W_i \in V_0^\perp(c)$. Using equation (10.3) of Definition 10.4, we obtain

$$I(W_i, W_i) = \int_{t=0}^{L} \sin^2 \frac{\pi t}{L} \left[g\left(R(E_i, c')c', E_i\right)\big|_t - \frac{\pi^2}{L^2} \right] dt.$$

Hence

$$\sum_{i=1}^{n-1} I(W_i, W_i) = \int_{t=0}^{L} \sin^2 \frac{\pi t}{L} \left[\mathrm{Ric}(c'(t), c'(t)) - \frac{(n-1)\pi^2}{L^2} \right] dt.$$

If $\mathrm{Ric}(c'(t), c'(t)) \geq (n-1)k$ for all $t \in [0, L]$ and $L \geq \pi/\sqrt{k}$, we find that $\sum_{i=1}^{n-1} I(W_i, W_i) \geq 0$. Hence $I(W_i, W_i) \geq 0$ for some $i \in \{1, 2, \ldots, n-1\}$. On the other hand, if $c \,|\, [0, L]$ is free of conjugate points, then $I(W_i, W_i) < 0$ for each i by Theorem 10.22. Hence c has a pair of conjugate points if $L \geq \pi/\sqrt{k}$, as required. □

A slight variant of Proposition 11.7 may also be proved similarly using the Timelike Morse Index Theorem (Theorem 10.27).

Proposition 11.8. *Let (M, g) be an arbitrary space–time of dimension n satisfying either (or both) of the curvature conditions of Proposition 11.7. If $c : [a, b] \to M$ is any timelike geodesic with $L(c) > \pi/\sqrt{k}$, then $t = a$ is conjugate along c to some $t_0 \in (a, b)$, and hence c is not maximal.*

Proof. Using $L(c) > \pi/\sqrt{k}$ and the same vector field W_i as in the proof of Proposition 11.7, we obtain this time that

$$\sum_{i=1}^{n-1} I(W_i, W_i) > 0.$$

Hence $I(W_i, W_i) > 0$ for some i. Thus $\operatorname{Ind}(c) > 0$. By the Timelike Morse Index Theorem (Theorem 10.27), $\dim J_t(c) \neq 0$ for some $t \in (a, b)$. This completes the proof. □

Now we are ready to give the Lorentzian analogue of the Bonnet–Myers Diameter Theorem for complete Riemannian manifolds.

Theorem 11.9. *Let (M, g) be a globally hyperbolic space–time of dimension n satisfying either of the following curvature conditions:*

(1) $K(\sigma) \leq -k < 0$ *for all timelike sectional curvatures* $K(\sigma)$.

(2) $\operatorname{Ric}(v, v) \geq (n-1)k > 0$ *for all unit timelike vectors* $v \in TM$.

Then $\operatorname{diam}(M, g) \leq \pi/\sqrt{k}$.

Proof. Suppose that $\operatorname{diam}(M, g) > \pi/\sqrt{k}$. We may then find $p, q \in M$ with $d(p, q) > \pi/\sqrt{k}$ by definition of $\operatorname{diam}(M, g)$. Since (M, g) is globally hyperbolic, there exists a maximal timelike geodesic segment $c : [0, 1] \to M$ with $c(0) = p$, $c(1) = q$. But as $L(c) = d(p, q) > \pi/\sqrt{k}$, the geodesic segment c is not maximal by Proposition 11.8, in contradiction. □

It is clear that an analogue of Proposition 11.8 could be obtained for null geodesics using the null index theory developed in Section 10.3. However, a stronger result may be obtained using the Raychaudhuri effect [cf. Hawking and Ellis (1973, p. 101)]. Thus we refer the reader to Section 12.2 for a discussion of these results rather than pursuing this analogy any further here [cf. also Harris (1982a)]. In addition to extensive results in the general relativity literature along the lines of Proposition 11.8, obtained using Raychaudhuri techniques, disconjugacy theory of O.D.E.'s has also been applied in this setting (cf. Tipler (1978), Galloway (1979), Kupeli (1986), just to single out a few of many possible references).

11.2 Lorentzian Comparison Theorems

For use in Section 11.3 as well as for their own intrinsic interest, we now present the timelike analogues of two important tools in global Riemannian geometry: the index comparison theorem [Gromoll, Klingenberg, and Meyer (1975, p. 174)] and the Rauch Comparison Theorem [Gromoll, Klingenberg, and Meyer (1975, p. 178) or Cheeger and Ebin (1975, p. 29)]. The results in this section, except for Corollary 11.12, have been published in Beem and Ehrlich (1979c, Section 9). Actually, there are two versions of the Rauch Comparison Theorem, often called *Rauch Theorem I* and *Rauch Theorem II*, that are useful in global Riemannian geometry [cf. Cheeger and Ebin (1975, Theorems 1.28 and 1.29, respectively)]. The result (Theorem 11.11) given in this section is the Lorentzian analogue of Rauch Theorem I. Harris (1979, 1982a) has given proofs of Lorentzian analogues for both Rauch Theorems I and II and using Rauch Theorem II has given a Lorentzian version of Toponogov's Comparison Theorem [cf. Cheeger and Ebin (1975, p. 42)] for timelike geodesic triangles in certain classes of space–times. Using this result, Harris (1979, 1982a) has also obtained a Lorentzian analogue of Toponogov's Diameter Theorem [cf. Cheeger and Ebin (1975, p. 110) and Appendix A].

In the rest of this section, let (M_1, g_1) and (M_2, g_2) be arbitrary space–times with $\dim M_1 \leq \dim M_2$. Also let $c_i : [0, L] \to M_i$, $i = 1, 2$, be *unit speed* future directed timelike geodesic segments. Throughout this section we will denote both the index form on $V^\perp(c_1)$ and $V^\perp(c_2)$ by I. Also, during the proofs we will denote both Lorentzian metrics g_1 and g_2 by $\langle\ ,\ \rangle$. The index form I for a timelike geodesic c and the index $\mathrm{Ind}(c)$ and extended index $\mathrm{Ind}_0(c)$ of c are defined in Section 10.1, equation (10.1) and Definition 10.24, respectively.

We first need to define an isomorphism

$$\phi : V^\perp(c_1) \to V^\perp(c_2)$$

so that $g_2(\phi X(t), \phi X(t)) = g_1(X(t), X(t))$ for all $t \in [0, L]$. This may be done following the usual parallel translation construction in Riemannian geometry. We first define an isometry

$$\phi_t : T_{c_1(t)}M_1 \to T_{c_2(t)}M_2$$

as follows. Let

$$P_t : T_{c_1(t)}M_1 \to T_{c_1(0)}M_1$$

denote the Lorentzian inner product-preserving isomorphism of parallel translation along c_1. Explicitly, given $v \in T_{c_1(t)}M_1$, let Y be the unique parallel field along c_1 with $Y(t) = v$, and set $P_t(v) = Y(0)$. Similarly, let

$$Q_t : T_{c_2(t)}M_2 \to T_{c_2(0)}M_2$$

denote parallel translation along c_2. Choose an injective Lorentzian inner product-preserving linear map

$$i : \left(T_{c_1(0)}M_1,\, g_1|_{c_1(0)}\right) \to \left(T_{c_2(0)}M_2,\, g_2|_{c_2(0)}\right)$$

where $i({c_1}'(0)) = {c_2}'(0)$. Then the map

$$\phi_t : \left(T_{c_1(t)}M_1,\, g_1|_{c_1(t)}\right) \to \left(T_{c_2(t)}M_2,\, g_2|_{c_2(t)}\right)$$

given by $\phi_t = Q_t^{-1} \circ i \circ P_t$ is an isometry since parallel translation preserves the Lorentzian structures. We may then define the map

$$\phi : V^\perp(c_1) \to V^\perp(c_2)$$

as follows. Given $X \in V^\perp(c_1)$, define $\phi X \in V^\perp(c_2)$ by $(\phi X)(t) = \phi_t(X(t))$. It follows as in the Riemannian proof that $(\phi X)' = \phi(X')$, where the first covariant differentiation is in M_2 and the second is in M_1.

Let $G_{2,t}(c_i)$ denote the set of all timelike planes σ containing ${c_i}'(t)$ for $i = 1, 2$. There is then an induced map

$$\phi_t : G_{2,t}(c_1) \to G_{2,t}(c_2)$$

defined as follows. If $\sigma \in G_{2,t}(c_1)$, we may write $\sigma = \mathrm{span}\{v, {c_1}'(t)\}$ where v is spacelike. Put $\phi_t(\sigma) = \mathrm{span}\{\phi_t(v), {c_2}'(t)\} \in G_{2,t}(c_2)$.

We are now ready to state the timelike version of the index comparison theorem.

Theorem 11.10 (Timelike Index Comparison Theorem). *Let (M_1, g_1) and (M_2, g_2) be space–times with $\dim M_1 \leq \dim M_2$, and let $c_1 : [0, \beta] \to M_1$ and $c_2 : [0, \beta] \to M_2$ be unit speed, future directed timelike geodesics. Suppose for all t with $0 \leq t \leq \beta$ and all timelike planes $\sigma \in G_{2,t}(c_1)$ that the sectional curvature $K_{M_1}(\sigma) \geq K_{M_2}(\phi_t \sigma)$. Then for any $X \in V^\perp(c_1)$ we have*

(1) $I(X, X) \leq I(\phi X, \phi X)$;
(2) $\text{Ind}(c_1) \leq \text{Ind}(c_2)$; *and*
(3) $\text{Ind}_0(c_1) \leq \text{Ind}_0(c_2)$.

Proof. Recalling that $\langle c_1', c_1' \rangle = -1$ and $\langle X, c_1' \rangle = 0$, we obtain

$$\langle R(X, c_1')c_1', X \rangle = -\langle X, X \rangle \, K(X, c_1').$$

Similar formulas hold for ϕX and c_2'. Thus

$$\begin{aligned} I(\phi X, \phi X) &= \int_0^\beta [-\langle (\phi X)', (\phi X)' \rangle + \langle R(\phi X, c_2')c_2', \phi X \rangle] \, dt \\ &= \int_0^\beta [-\langle \phi(X'), \phi(X') \rangle + \langle R(\phi X, c_2')c_2', \phi X \rangle] \, dt \\ &\geq \int_0^\beta [-\langle X', X' \rangle + \langle R(X, c_1')c_1', X \rangle] \, dt = I(X, X). \quad \square \end{aligned}$$

We may now obtain the more powerful Timelike Rauch Comparison Theorem using the timelike index comparison theorem. Recall first that since $c_i : [0, L] \to M_i$ is a timelike geodesic segment for each i, the vector fields in $V^\perp(c_i)$ are all spacelike vector fields.

Theorem 11.11 (Timelike Rauch Comparison Theorem). *Let (M_1, g_1) and (M_2, g_2) be space–times with $\dim M_1 \leq \dim M_2$. Let $c_1 : [0, L] \to M_1$ and $c_2 : [0, L] \to M_2$ be future directed, timelike, unit speed geodesic segments. Suppose for all $t \in [0, L]$ and any $\sigma \in G_{2,t}(c_1)$ that*

$$K_{M_1}(\sigma) \geq K_{M_2}(\phi_t \sigma).$$

Let $Y_1 \in V^\perp(c_1)$ and $Y_2 \in V^\perp(c_2)$ be Jacobi fields on M_1 and M_2 respectively, satisfying the initial conditions

$$Y_1(0) = Y_2(0) = 0, \tag{11.1}$$

and

$$g_1(Y_1'(0), Y_1'(0)) = g_2(Y_2'(0), Y_2'(0)). \tag{11.2}$$

If c_2 has no conjugate points to $t = 0$ in $(0, L)$, then

$$g_1(Y_1(t), Y_1(t)) \geq g_2(Y_2(t), Y_2(t)) \tag{11.3}$$

for all $t \in (0, L]$. In particular, c_1 has no conjugate points to $t = 0$ in $(0, L)$.

Proof. This may be given along the lines of Gromoll, Klingenberg, and Meyer (1975, pp. 180–181), except for the proof of their inequality (7), p. 181, which must be modified as follows. Let $Z \in V^\perp(c_2)$ be the unique Jacobi field along c_2 with $Z(0) = 0$ and $Z(t_0) = \phi Y_1(t_0)$, ϕ as above. We must show that

$$\langle Y_1(t_0), Y_1'(t_0)\rangle \geq \langle Z(t_0), Z'(t_0)\rangle.$$

To this end, let $c_3 = c_1 \mid [0, t_0]$ and $c_4 = c_2 \mid [0, t_0]$. Then

$$\begin{aligned}\langle Y_1(t_0), Y_1'(t_0)\rangle &= -I(Y_1 \mid c_3,\ Y_1 \mid c_3)\\ &\geq -I((\phi Y_1) \mid c_4,\ (\phi Y_1) \mid c_4)\end{aligned}$$

(the above inequality by the timelike index comparison theorem [Theorem 11.10–(1)])

$$\geq -I(Z \mid c_4, Z \mid c_4)$$

(the above inequality by the maximality of Jacobi fields with respect to the index form in the absence of conjugate points [Theorem 10.23])

$$= \langle Z(t_0), Z'(t_0)\rangle. \quad \square$$

It may also be assumed in Theorem 11.11 that c_2 has no conjugate points in $[0, L]$. Then c_1 would also have no conjugate points in $[0, L]$ by Theorem 11.10–(3).

For Riemannian manifolds of nonpositive sectional curvature, by equipping the tangent space with the "flat metric" (cf. Definition 10.17) it may be shown that the exponential map does not decrease the length of tangent vectors [cf. Bishop and Crittenden (1964, p. 178, Theorem 2 (i)) for a precise statement]. A simple proof of this fact may be given using the Rauch Comparison Theorem and comparing Jacobi fields on the given Riemannian manifold to those in $\mathbb{R}^n$. We will now use the Timelike Rauch Comparison Theorem to prove the analogous result for space–times of nonnegative timelike sectional curvature [cf. Flaherty (1975a, p. 397)]. Intuitively, Corollary 11.12 below expresses the fact that if all timelike sectional curvatures of (M, g) are positive, then future directed timelike geodesics emanating from a given point of M spread apart faster than "corresponding" geodesics in Minkowski space–time. Recall that the canonical isomorphism τ_v has been defined in Section 10.1, Definition 10.15.

Corollary 11.12. *Let (M, g) be a space–time with everywhere nonnegative timelike sectional curvature, and let $v \in T_pM$ be a given future directed timelike tangent vector with $g(v, v) = -1$. Then for any future directed nonspacelike tangent vector $w \in T_pM$, the vector $b = \tau_v(w) \in T_v(T_pM)$ satisfies the inequality*

$$g(\exp_{p_*} b, \exp_{p_*} b) \geq g(w, w) = \langle\langle b, b \rangle\rangle.$$

Proof. We first prove the inequality for $b = \tau_v(w) \in T_v(T_pM)$ with $g(v, w) = 0$ by applying the Timelike Rauch Comparison Theorem with $(M_1, g_1) = (M, g)$ and (M_2, g_2) identified with Minkowski space–time $(\mathbb{R}^n_1, g_0)$ with $n = \dim M$. Setting $c_1(t) = \exp_p tv$, let $Y_1 \in V^\perp(c)$ be the unique Jacobi field with $Y_1(0) = 0$, $Y_1'(0) = w$. By Proposition 10.16, we have $Y_1(1) = \exp_{p_*} b$.

Now let $c_2 : [0, 1] \to \mathbb{R}^n_1$ be an arbitrary unit speed timelike geodesic, and choose $\overline{w} \in N(c_2(0))$ with $g_0(\overline{w}, \overline{w}) = g_1(w, w)$. Let $Y_2 \in V^\perp(c_2)$ be the unique Jacobi field with $Y_2(0) = 0$ and $Y_2'(0) = \overline{w}$. Then $Y_2(t) = tP_t(\overline{w})$, where P_t denotes the Lorentzian parallel translation along c_2 from $c_2(0)$ to $c_2(t)$. Applying Theorem 11.11, we obtain

$$\begin{aligned} g_1(\exp_{p_*} b, \exp_{p_*} b) &= g_1(Y_1(1), Y_1(1)) \\ &\geq g_0(Y_2(1), Y_2(1)) \\ &= g_0(P_1(\overline{w}), P_1(\overline{w})) \\ &= g_0(\overline{w}, \overline{w}) = g_1(w, w) = \langle\langle b, b \rangle\rangle \end{aligned}$$

as required.

Returning to the general case, we may decompose $w = w_1 + w_2$ where $w_1 = \lambda v$ for some $\lambda > 0$ and $g(v, w_2) = 0$. Set $b_i = \tau_v(w_i)$ for $i = 1, 2$ so that $b = b_1 + b_2$. We now calculate

$$\begin{aligned} g(\exp_{p_*} b, \exp_{p_*} b) &= g(\exp_{p_*} b_1, \exp_{p_*} b_1) \\ &\quad + 2\, g(\exp_{p_*} b_1, \exp_{p_*} b_2) + g(\exp_{p_*} b_2, \exp_{p_*} b_2). \end{aligned}$$

Applying the Gauss Lemma (Theorem 10.18) to the first two terms, we obtain

$$\begin{aligned} g(\exp_{p_*} b, \exp_{p_*} b) &= \langle\langle b_1, b_1 \rangle\rangle + 2\langle\langle b_1, b_2 \rangle\rangle + g(\exp_{p_*} b_2, \exp_{p_*} b_2) \\ &= \langle\langle b_1, b_1 \rangle\rangle + g(\exp_{p_*} b_2, \exp_{p_*} b_2) \end{aligned}$$

as $\langle\langle b_1, b_2 \rangle\rangle = g(w_1, w_2) = 0$. Now applying the first part of the proof to the last term, we have

$$\begin{aligned} g(\exp_{p_*} b, \exp_{p_*} b) &\geq \langle\langle b_1, b_1 \rangle\rangle + \langle\langle b_2, b_2 \rangle\rangle \\ &= \langle\langle b_1 + b_2, b_1 + b_2 \rangle\rangle = \langle\langle b, b \rangle\rangle \end{aligned}$$

as required. □

11.3 Lorentzian Hadamard–Cartan Theorems

We first prove a basic result linking conjugate points and timelike sectional curvature [cf. Flaherty (1975a, Proposition 2.1)—here Flaherty uses the signature convention $(+, -, \cdots, -)$ for the Lorentzian metric so that his sectional curvature condition has the opposite sign from ours].

Proposition 11.13. *Let (M, g) be a space–time with everywhere nonnegative timelike sectional curvatures. Then no nonspacelike geodesic has any conjugate points.*

Proof. First let $c : [0, a) \to M$ be an arbitrary future directed unit speed timelike geodesic. Recall from Corollary 10.10 that if X is a Jacobi field along c with $X(0) = X(t_0) = 0$ for some $t_0 \in (0, a)$, then $X \in V^\perp(c)$. Thus we may restrict our attention to Jacobi fields $J \in V^\perp(c)$ with $J(0) = 0$. Since $J \in V^\perp(c)$ and c is a timelike geodesic, $J' \in V^\perp(c)$ also. Consider the smooth function $f(t) = g(J(t), J'(t))$. Differentiating yields

$$\begin{aligned} f'(t) &= g(J'(t), J'(t)) + g(J(t), J''(t)) \\ &= g(J'(t), J'(t)) - g\left(J(t),\, R(J(t), c'(t))c'(t)\right) \\ &= g(J'(t), J'(t)) + g\left(J(t), J(t)\right) K(J(t), c'(t)) \geq 0 \end{aligned}$$

for all $t \in [0, a)$. Now if $J(t_0) = 0$ for some $t_0 \in [0, a)$, then $f(0) = f(t_0) = 0$. Thus as f is nondecreasing, $f(t) = 0$ for all $t \in [0, t_0]$. Hence $0 = f'(0) = g(J'(0), J'(0))$ as $J(0) = 0$, from which we conclude that $J'(0) = 0$. Therefore $J = 0$, and no $t_0 \in [0, a)$ is conjugate to $t = 0$ along c.

We now treat the case that $\beta : [0, a) \to M$ is a null geodesic using the null index form. Let $t_0 \in [0, a)$ be arbitrary. We will show that $\overline{I} : \mathfrak{X}_0(\beta \,|\, [0, t_0]) \times \mathfrak{X}_0(\beta \,|\, [0, t_0]) \to \mathbb{R}$ is negative definite. Hence no $s \in (0, t_0]$ is conjugate to $t = 0$ along β by Theorem 10.69.

If $W \in \mathfrak{X}_0(\beta)$ is a smooth parallel vector class, then $W = [\beta']$ since $W(0) = [\beta'(0)]$. Thus if $W \in \mathfrak{X}_0(\beta)$ is not a smooth parallel vector class, we have $\overline{g}(W'(s), W'(s)) > 0$ for some $s \in (0, t_0)$. Also $\overline{g}(W'(t), W'(t)) \geq 0$ for all $t \in [0, t_0]$ since $\overline{g}$ is positive definite. Now consider $\overline{g}\left(\overline{R}(W(t), \beta'(t))\beta'(t), W(t)\right)$ for any fixed $t \in [0, t_0]$. If $W(t) = [\beta'(t)]$, then $\overline{g}\left(\overline{R}(W(t), \beta'(t))\beta'(t), W(t)\right) = 0$. Otherwise, we may find a spacelike tangent vector w perpendicular to $\beta'(t)$ with $\pi(w) = W(t)$. Then, using equation (10.35) of Section 10.3 we have $\overline{g}\left(\overline{R}(W(t), \beta'(t))\beta'(t), W(t)\right) = g\left(R(w, \beta'(t))\beta'(t), w\right)$. Since w is spacelike, we may find a sequence of timelike two-planes $\sigma_n = \{w_n, v_n\}$ with w_n spacelike, v_n timelike, $g(w_n, v_n) = 0$, $w_n \to w$, and $v_n \to \beta'(t)$, so that $\sigma_n \to \{w, \beta'(t)\}$.

We then obtain by continuity

$$\begin{aligned}\overline{g}\left(\overline{R}(W(t),\beta'(t))\beta'(t),W(t)\right) &= g\left(R(w,\beta'(t))\beta'(t),w\right)\\ &= \lim_{n\to\infty} g(R(w_n,v_n)v_n,w_n)\\ &= \lim_{n\to\infty} K(w_n,v_n)\,g(w_n,w_n)\,g(v_n,v_n) \le 0\end{aligned}$$

since $g(w_n,w_n) > 0$, $g(v_n,v_n) < 0$, and $K(v_n,w_n) \ge 0$ for each n. Thus in either case, $\overline{g}\left(\overline{R}(W(t),\beta'(t))\beta'(t),W(t)\right) \le 0$. Hence, provided $W \ne [\beta']$ we have

$$\overline{I}(W,W) = \int_{t=0}^{t_0} \left[-\overline{g}(W',W') + \overline{g}(\overline{R}(W,\beta')\beta',W)\right]\Big|_t\, dt < 0$$

as required. □

Motivated by a standard definition in global Riemannian geometry, we make the following definition.

Definition 11.14. The space–time (M,g) is said to have *no future timelike conjugate points* if for any future directed timelike geodesic $c:[0,a)\to(M,g)$, no nontrivial Jacobi field in $V^{\perp}(c)$ vanishes more than once.

In view of Lemma 10.46, similar definitions may be formulated for space–times with no future null conjugate points or no future nonspacelike conjugate points. Proposition 11.13 guarantees that if (M,g) is a space–time with everywhere nonnegative timelike sectional curvature, then (M,g) has no future nonspacelike conjugate points.

Lorentzian manifolds with nonnegative timelike sectional curvature or with no future timelike conjugate points may be characterized in terms of the behavior of their Jacobi fields. A similar characterization applies to Riemannian manifolds [cf. O'Sullivan (1974, Proposition 4)].

Proposition 11.15.

(1) *(M,g) has everywhere nonnegative timelike sectional curvature iff*

$$\frac{d^2}{dt^2}(g(Y(t),Y(t))) \ge 0$$

for every Jacobi field $Y \in V^{\perp}(c)$ along any future directed timelike geodesic c.

(2) *(M, g) has no future timelike conjugate points iff $g(Y(t), Y(t)) > 0$ for all $t > 0$ where $Y \in V^\perp(c)$ is any nontrivial Jacobi field with $Y(0) = 0$ along any future directed timelike geodesic c.*

Proof. (1) Assume that (M, g) has everywhere nonnegative timelike sectional curvature. Let $Y \in V^\perp(c)$ be a Jacobi field along the unit speed timelike geodesic c. Then $Y' \in V^\perp(c)$ also, and we obtain

$$\begin{aligned}\frac{d^2}{dt^2}(g(Y,Y)) &= 2\,g(Y',Y') - 2\,g(R(Y,c')c',Y)\\ &= 2\,g(Y',Y') + 2\,g(Y,Y)\,K(Y,c') \geq 0.\end{aligned}$$

Conversely, let $\{v, w\}$ be future timelike and spacelike tangent vectors, respectively, spanning an arbitrary timelike two-plane with $g(v, v) = -1$, $g(w, w) = 1$, and $g(v, w) = 0$. Let $c(t) = \exp(tv)$ and let $Y \in V^\perp(c)$ be the Jacobi field with initial conditions $Y(0) = w$ and $Y'(0) = 0$. Then we have by hypothesis,

$$\begin{aligned}0 \leq \frac{d^2}{dt^2}(g(Y,Y))\bigg|_{t=0} &= -2\,g\left(R(Y(0),c'(0))c'(0),Y(0)\right)\\ &= -2\,g\left(R(w,v)v,w\right) = 2K(v,w)\end{aligned}$$

since the first term in the differentiation vanishes as $Y'(0) = 0$. Thus $K(v, w) \geq 0$ as required.

(2) This is clear from Definition 11.14. □

Using the timelike index theory of Sections 10.1 and 10.2, we now give the following version of a Lorentzian Hadamard–Cartan Theorem for globally hyperbolic space–times. This proof is similar to the Morse theory proof of the Hadamard–Cartan Theorem for complete Riemannian manifolds [cf. Milnor (1963, p. 102)]. Recall that a space–time is said to be *future one-connected* (Definition 10.28) if any two smooth, future directed timelike curves from p to q are homotopic through (smooth) future directed timelike curves with fixed endpoints p and q.

Theorem 11.16. *Let (M, g) be a future one-connected globally hyperbolic space–time with no future nonspacelike conjugate points. Then given any*

$p, q \in M$ with $p \ll q$, there is exactly one future directed timelike geodesic (up to reparametrization) from p to q.

Proof. Since (M, g) is globally hyperbolic, there exists a maximal future directed timelike geodesic from p to q. Since there are no future nonspacelike conjugate points, any future directed geodesic from p to q has index 0 by Theorem 10.27. Thus the timelike path space $C_{(p,q)}$ has the homotopy type of a CW-complex with a cell of dimension 0, i.e., a point, for each future directed timelike geodesic from p to q. On the other hand, since M is future one-connected, $C_{(p,q)}$ is connected and hence consists of a single point. Thus there is at most one future directed timelike geodesic from p to q. $\square$

A similar result was obtained by Uhlenbeck (1975, Theorem 5.3) for globally hyperbolic space–times satisfying a metric growth condition [Uhlenbeck (1975, p. 72)] and the curvature condition $g(R(v,w)w, v) \leq 0$ for all future directed null vectors v and vectors w with $g(v, w) = 0$ at every point of M. Namely, M can be covered by a space which is topologically Minkowski (i.e., Euclidean) space.

Flaherty (1975a, p. 398) has also shown that if (M, g) is future 1-connected, future nonspacelike complete, and has everywhere nonnegative timelike sectional curvatures, then the exponential map $\exp_p$ regularly embeds the future cone in T_pM at each point p into M. To obtain this result, Flaherty used a lifting argument to show that under these hypotheses, if $v, w \in T_pM$ are any two future directed timelike tangent vectors with $\exp_p v = \exp_p w$, then $v = w$. Thus future one-connected, future nonspacelike complete space–times with nonnegative timelike sectional curvatures satisfy the conclusion of Theorem 11.16. On the other hand, Flaherty showed (1975b, p. 200) that any future one-connected, future nonspacelike complete space–time with everywhere nonnegative timelike sectional curvatures is also globally hyperbolic [cf. Galloway (1986a, 1989b)].

Recall that in Section 7.4 we applied disprisonment and pseudoconvexity to nonspacelike geodesics to obtain sufficient conditions for stability of nonspacelike geodesic incompleteness (nonspacelike disprisonment suffices) and the stability of nonspacelike geodesic completeness (nonspacelike disprisonment and nonspacelike pseudoconvexity taken together suffice). In order to get a Hadamard–Cartan Theorem that applies to the entire space–time, we will apply these two conditions to the set of all geodesics—not just to nonspacelike geodesics [cf. Beem and Parker (1989)]. A manifold will be geodesically disprisoning if each end of every inextendible geodesic fails to be imprisoned in a compact set. The manifold will be geodesically pseudoconvex if for each compact set K the convex hull, using geodesic segments joining points of K, lies in a larger compact set. More precisely,

Definition 11.17. (*Geodesically Disprisoning*) The space–time (M, g) is *geodesically disprisoning* if, for each inextendible geodesic $c\colon (a, b) \to M$ and any fixed $t_0 \in (a, b)$, the images of each of the two maps $c \mid (a, t_0]$ and $c \mid [t_0, b)$ fail to have compact closure.

Definition 11.18. (*Geodesically Pseudoconvex*) The space–time (M, g) is *geodesically pseudoconvex* if for each compact subset K of M there is a compact set H such that each geodesic segment $c : [a, b] \to M$ with $c(a), c(b) \in K$ satisfies $c([a, b]) \subseteq H$.

By itself, Definition 11.17 allows for the possibility that either (or both) ends of a given geodesic may be partially imprisoned (cf. Definition 7.29). However, the next lemma shows that *if (M, g) satisfies both of the above conditions, then partial imprisonment cannot occur*.

Lemma 11.19. *Let (M, g) be both geodesically disprisoning and geodesically pseudoconvex. If $c\colon (a, b) \to M$ is any inextendible geodesic of M and K is any compact subset of M, then there are parameter values s_1, s_2 with $a < s_1 < s_2 < b$ such that $c(t) \notin K$ for all $a < t < s_1$ and all $s_2 < t < b$.*

Proof. Assume the lemma is false. Then without loss of generality we may assume there is some geodesic $c\colon (a, b) \to M$ and a sequence $\{t_k\}$ satisfying $a < t_1 < \cdots < t_k < \cdots < b$, $t_k \to b^-$, and $c(t_k) \in K$ for some compact subset K of M. Let H be a compact set such that any geodesic segment of (M, g) with endpoints in K lies in H. Then the image of the map $c \mid [t_1, b)$ lies in the compact set H, in contradiction to the disprisonment assumption. □

Lemma 11.20. *Let (M, g) be both geodesically disprisoning and geodesically pseudoconvex. Let $\{p_n\}$ and $\{r_n\}$ be sequences converging to p and r, respectively. If each pair $\{p_n, r_n\}$ is joined by a geodesic c_n, then there is a geodesic c joining p and r. Furthermore, if each c_n is nonspacelike (respectively, null), then c may be chosen to be nonspacelike (respectively, null).*

Proof. Let h be an auxiliary Riemannian metric on M. Let K be a compact set containing p and r as well as all $\{p_n\}$ and $\{r_n\}$. For each n, let $c_n\colon [0, b_n) \to M$ be a geodesic for the metric g which satisfies $c_n(0) = p_n$, $c_n(a_n) = r_n$, and which is not extendible to b_n. We may assume without loss of generality that $h(c_n'(0), c_n'(0)) = 1$. Since h is positive definite, the sphere bundle over K with respect to h is compact. Consequently, there is a subsequence $\{m\}$ of $\{n\}$ and a vector $w \in T_pM$ with $c_m'(0) \to w$. Let $c\colon [0, b) \to M$ be the unique geodesic with $c'(0) = w$ and such that c is not extendible to $t = b$. Let H be a compact set such that all geodesic segments with endpoints in K lie in H. Since (M, g) is disprisoning, there is some $0 < \tau < b$ with $c(\tau) \notin H$. The construction of H then implies $c(t) \notin K$ for all $\tau \leq t < b$. Since $c_m(\tau) \to c(\tau)$, it follows that $0 < a_m < \tau$ for all sufficiently large m. Using the fact that $[0, \tau]$ is compact, we find there is a limit point a of the sequence $\{a_m\}$. Let $\{k\}$ be a subsequence of $\{m\}$ with $a_k \to a$. Then $c_k(a_k) \to c(a)$, and it follows that $c(a) = r = \lim r_k$. Hence c is a geodesic containing both p and r. If each c_n is nonspacelike (respectively, null), then $c'(0) = \lim\, c_k'(0)$ shows c is also nonspacelike (respectively, null). □

Definition 11.21. (*Geodesically Connected*) The space–time (M, g) is *geodesically connected* if for each pair of distinct points p and q there is at least one geodesic segment joining p to q.

The Hopf–Rinow Theorem [cf. Hopf and Rinow (1931)] guarantees the geodesic connectedness of any complete Riemannian manifold. However, complete Lorentzian manifolds may fail to be geodesically connected. In fact, for complete Lorentzian manifolds, even if r is in the causal future of p, it may happen that there is no geodesic from p to r (cf. Figure 6.1). Of course, global hyperbolicity yields a (maximal length) geodesic segment joining any two causally related points. However, geodesic completeness does not imply global hyperbolicity. Furthermore, global hyperbolicity does not yield the existence of geodesic segments between points that fail to be causally related. The next proposition gives sufficient conditions for the existence of a geodesic segment between any two points. The hypotheses include disprisonment, pseudoconvexity, and a lack of conjugate points but do not require geodesic completeness.

Proposition 11.22. *Let (M, g) be geodesically disprisoning and geodesically pseudoconvex. If (M, g) has no conjugate points, then (M, g) is geodesically connected.*

Proof. Let p and r be distinct points of M, and construct an arbitrary curve $\alpha\colon [0,1] \to M$ with $\alpha(0) = p$ and $\alpha(1) = r$. Let $\tau = \sup\{s \in [0,1] :$ there is a geodesic segment from p to $\alpha(t)$ for all $0 \le t \le s\}$. The fact that p lies in a convex normal neighborhood yields $\tau > 0$. Lemma 11.20 implies that there is a geodesic segment $c\colon [0,1] \to M$ with $c(0) = p$ and $c(1) = \alpha(\tau)$. Let $X = c'(0)$. Since (M, g) has no conjugate points, the exponential map $\exp_p$ takes a sufficiently small open neighborhood of X onto an open neighborhood of $c(1) = \alpha(\tau)$. Thus, there exist geodesic segments from p to all points of α sufficiently near $\alpha(\tau)$. It follows that $\tau = 1$, and hence there is a geodesic segment from p to r, as desired. □

Compact Riemannian manifolds are always geodesically connected. In contrast, the next example shows that compact Lorentzian manifolds may fail to be geodesically connected.

Example 11.23. (*Compact space–times need not be geodesically connected*) Let $M = S^1 \times S^1 = \{(t, s) : 0 \le s \le 1 \text{ and } -4\pi \le t \le 4\pi\}$ with the usual identifications. Let $g = (-\cos t)dt^2 + (2\sin t)dtds + (\cos t)ds^2$. This is a Lorentzian metric, and $X = \cos(t/2)\,\partial/\partial t - \sin(t/2)\,\partial/\partial s$ is a unit timelike vector field which time orients M. Using light cones, it is not hard to check that all geodesics starting on the circle $t = 0$ must lie in the set $-5\pi/2 \le t \le 7\pi/2$. It follows that (M, g) is a compact space–time which is not geodesically connected.

Let G be a group of homeomorphisms acting on the manifold M. The group G is said to act *freely* if $f(p) = p$ for some $p \in M$ implies that f is the identity e of G. If G acts freely, then e is the only element of G that has any fixed points. A discrete group G is said to act *properly discontinuously* [cf. Boothby (1986), Wolf (1974)] if the following hold:

(1) Each $p \in M$ has a neighborhood U such that the set $\{f \in G : f(U) \cap U \neq \emptyset\}$ is finite; and

(2) If $f(p) \neq q$ for all $f \in G$, then there are neighborhoods V and W of p and q, respectively, such that $f(V) \cap W = \emptyset$ for all $f \in G$.

Group actions arise naturally in the study of covering spaces of manifolds. Let M be a smooth manifold. A *covering space* of M is a triple (E, π, M) where E is a smooth manifold and $\pi \colon E \to M$ is a smooth map of E onto M which is evenly covered [cf. O'Neill (1983)]. Here π is said to be *evenly covered* if for each $p \in M$ there is a neighborhood U of p such that $\pi^{-1}(U)$ consists of disjoint open sets $\{W_\alpha\}$ in E and $\pi \mid W_\alpha$ is a diffeomorphism of W_α onto U for each α. If E is simply connected, then (E, π, M) is the *universal covering space* of M. A covering map π is a local diffeomorphism, but not all local diffeomorphisms are covering maps because a local diffeomorphism is not necessarily an even cover. For example, let $M = S^1 = \{(x, y) : x^2 + y^2 = 1\} = \{z \in \mathbb{C} : |z| = 1\}$. The map $f \colon (0, 4\pi) \to S^1$ defined by $f(\theta) = e^{i\theta}$ is a local diffeomorphism but fails to be a covering map. The inverse image of $e^{i0} = (1, 0)$ is the single number 2π. Other points have two preimages, and the inverse image of a small connected open set U about the point $(1, 0)$ in S^1 is three open intervals, one of which contains the number 2π. The map f restricted to the interval of $f^{-1}(U)$ containing 2π is a diffeomorphism onto U, but f restricted to either of the other two intervals is not a diffeomorphism onto U.

The following topological result will be of use in the proof of Theorem 11.25.

Lemma 11.24. *If $\pi : \mathbb{R}^n \to M$ is the universal covering space of the manifold M, then the fundamental group $\pi_1(M)$ of M is either trivial or infinite. If $\pi_1(M)$ is trivial, then the covering map π is one-to-one.*

Proof. The fundamental group $\pi_1(M)$ of M acts freely and properly discontinuously on the total space $\mathbb{R}^n$ of the covering [cf. Kobayashi and Nomizu

(1963, p. 61)]. If $\pi_1(M)$ is finite, then it is either trivial or else has an element of prime order. However, a result of Smith (1941, p. 3) yields that any homeomorphism of $\mathbb{R}^n$ having prime order must fix some point of $\mathbb{R}^n$. But this would contradict the free action of $\pi_1(M)$. Thus, the fundamental group is infinite or has exactly one element. If the fundamental group is trivial, then each $p \in M$ has exactly one preimage and hence the covering is one-to-one. □

The next theorem is a Lorentzian version of the Hadamard–Cartan Theorem and is actually valid for any (connected) manifold with an affine connection [cf. Beem and Parker (1989)]. The usual completeness assumption used in the Riemannian version [cf. Kobayashi (1961)] is replaced by disprisonment and pseudoconvexity assumptions. In some sense, the pseudoconvexity condition is a type of "internal completeness" assumption which plays a role similar to global hyperbolicity (cf. Section 7.4). Since global hyperbolicity is a causality assumption, it only provides a certain amount of "control" over nonspacelike geodesics. It should also be recalled that global hyperbolicity, which was used in the hypothesis of our previous version of a Lorentzian Hadamard–Cartan Theorem, Theorem 11.16 earlier in this section, implies *nonspacelike* geodesic pseudoconvexity (cf. Proposition 7.36).

Theorem 11.25. *Let (M, g) be geodesically disprisoning and geodesically pseudoconvex. If (M, g) has no conjugate points, then for each point $p \in M$, the exponential map $\exp_p : D \to M$ is a diffeomorphism from the domain $D \subseteq T_pM$ of $\exp_p$ onto M. Consequently, M is diffeomorphic to $\mathbb{R}^n$, M is geodesically connected, and any two distinct points of M determine exactly one geodesic.*

Proof. The map $\exp_p$ is onto M by Proposition 11.22. Furthermore, the exponential map at p must be at least a local diffeomorphism from its domain D onto M since no conjugate points exist. In order to show that $\exp_p$ is a diffeomorphism, we will first show that each $r \in M$ has a finite number of preimages. Secondly, the map $\exp_p$ will be shown to be a covering map, and finally, Lemma 11.24 will yield that $\exp_p$ is one-to-one.

To show that a fixed point r of M has a finite number of preimages in D, we begin by choosing a compact set K of M containing p and with r in the interior of K. Then there is an open neighborhood $U(r)$ of r with $r \in U(r) \subseteq K$. Let H be a compact set such that any geodesic segment with endpoints in K lies in H. Pseudoconvexity guarantees the existence of H. Choose an auxiliary Riemannian metric h, and let $S_h(p) = \{v \in T_pM : h(v,v) = 1\}$. Let $v \in S_h(p)$ be arbitrary. If $c_v : [0,b) \to M$ is the geodesic of g with $c_v'(0) = v$, then disprisonment yields $c_v(\tau) \notin H$ for some τ. It follows that $c_v(t) \notin K$ for all $\tau \leq t < b$. Using the continuous dependence of geodesics on the initial tangent vector, it follows that there is a neighborhood $N(v)$ of v in $S_h(p)$ such that each geodesic with initial tangent vector lying in this neighborhood must have τ in its domain and must have image points not in K for all $t > \tau$. In particular, if c_v happens to intersect $U(r)$, then the intersection must occur for parameter values not greater than τ. Let $\Gamma(v) = \{w \in T_pM : w = \alpha v$ with $v \in S_h(p)$ and $0 \leq \alpha \leq \tau\}$. The set $\Gamma(v)$ is compact and lies in the domain of the local diffeomorphism $\exp_p$. Consequently, each point $q \in U(r)$ has only a finite number of preimages in $\Gamma(v)$. We cover the compact set $S_h(p)$ with a finite number of neighborhoods $N(v_1), \ldots, N(v_m)$ of the above type for corresponding vectors $v_1, \ldots, v_m$. Thus, we obtain a finite number of compact sets $\Gamma(v_1), \ldots, \Gamma(v_m)$ each having a finite number of preimages of any fixed point $q \in U(r)$. Let $\Gamma = \Gamma(v_1) \cup \cdots \cup \Gamma(v_m)$. By construction, all preimages of each $q \in U(r)$ lie in the compact set Γ. Hence we find that $r \in M$ has a finite number of preimages in $D \subseteq T_pM$.

Assume that $\exp_p$ is not a covering map. There must be some $r \in M$ such that for each open neighborhood U of r, the set $\exp_p^{-1}(U)$ fails to consist of components each of which is diffeomorphic to U under $\exp_p$. Let K, $U(r)$, H, and Γ be as above. Let r have k preimages $u_1, \ldots, u_k$. Since $\exp_p$ is a local diffeomorphism, there is an open neighborhood $V \subseteq U(r)$ of r diffeomorphic to an open ball, and corresponding disjoint open neighborhoods $W_1, \ldots, W_k$ in T_pM containing the individual preimages of r, such that $\exp_p \mid W_i$ is a diffeomorphism onto V for each i. It follows that $\exp_p^{-1}(V)$ consists of more than $W_1 \cup W_2 \cup \cdots \cup W_k$, and this must be true no matter how small V is chosen. Consequently, there is a sequence $r_n \to r$ such that the number of preimages of each r_n is larger than k. This yields a sequence $\{w_n\}$ in T_pM

with $\exp_p(w_n) = r_n$ such that no subsequence of $\{w_n\}$ converges to one of the preimages $u_1, \ldots, u_k$ of r. Since each w_n lies in the compact set Γ, there is a subsequence $\{w_j\}$ converging to some w in Γ. Thus, w is in the domain of $\exp_p$, and $\exp_p(w) = \lim \exp_p(w_j) = \lim r_j = r$, which shows w must be one of the k preimages of r, in contradiction. We conclude that $\exp_p$ must be a covering map. The (star shaped) domain D of $\exp_p$ is some open set in T_pM, and D must be homeomorphic to $\mathbb{R}^n$. Thus $\exp_p : D \to M$ is a universal covering and the number of elements in $\pi_1(M)$ is equal to the number of preimages of any point $r \in M$. Lemma 11.24 now yields that $\pi_1(M)$ is trivial and hence that $\exp_p : D \to M$ is a diffeomorphism, as desired. □

Let a Lorentzian manifold (M, g) be given, and assume $\pi : \overline{M} \to M$ is a covering of M by $\overline{M}$. One may obtain an induced metric $\overline{g}$ on $\overline{M}$ such that π is a local isometry [cf. Wolf (1974), pp. 41–42]. The geodesics of $\overline{M}$ project to geodesics of M, and geodesics of M lift to geodesics of $\overline{M}$. It may easily happen that the cover is geodesically disprisoning even if the original base manifold (M, g) fails to be disprisoning. For example, one could have $M = S^1 \times S^1 = \{(t, x) : 0 \leq t \leq 1,\ 0 \leq x \leq 1\}$ with the usual identifications, given the flat metric $g = -dt^2 + dx^2$. Then $(\overline{M}, \overline{g})$ is the usual two-dimensional Minkowski space–time.

Clearly, Theorem 11.25 may be applied to the cover to obtain information on the base manifold M.

Corollary 11.26. *Let (M, g) have a covering space $\pi : \overline{M} \to M$ with induced metric $\overline{g} = \pi^* g$ on $\overline{M}$. If this covering space $(\overline{M}, \overline{g})$ has no conjugate points, is geodesically pseudoconvex, and is geodesically disprisoning, then both $(\overline{M}, \overline{g})$ and (M, g) are geodesically connected. Furthermore, if p is any point of M, then the exponential map $\exp_p$ is a covering map.*

Proof. Theorem 11.25 yields the geodesic connectedness of the cover. To show that M is geodesically connected, first let p and q be two points of M. Lift them to points $\overline{p}$ and $\overline{q}$, respectively, in $\overline{M}$. Then the geodesic segment from $\overline{p}$ to $\overline{q}$ projects to a geodesic segment from p to q. Thus the geodesic connectivity of the cover implies the geodesic connectivity of the base.

It remains to show that $\exp_p$ is a covering map. Select a point $\overline{p} \in \pi^{-1}(p)$. Let $D_p \subseteq T_pM$ be the domain of $\exp_p$ and $D_{\overline{p}} \subseteq T_{\overline{p}}\overline{M}$ be the domain of $\exp_{\overline{p}}$. The composition $\pi \circ \exp_{\overline{p}} : D_{\overline{p}} \to M$ is a covering map since $\exp_{\overline{p}}$ is a diffeomorphism onto $\overline{M}$ and π is a covering map with domain $\overline{M}$. Recall that π is an isometry of any sufficiently small neighborhood W of $\overline{p}$ to a corresponding neighborhood U of p. Also, each geodesic through $\overline{p}$ projects to a geodesic through p, and each geodesic though p lifts (uniquely) to a geodesic through $\overline{p}$. It follows that the differential of the map $(\pi \,|\, W)^{-1}$ at p yields a diffeomorphism $f : D_p \to D_{\overline{p}}$ such that $\exp_p = \pi \circ \exp_{\overline{p}} \circ f$. Consequently, $\exp_p : D_p \to M$ is a covering map, as desired. □

Versions of Proposition 11.22 and Theorem 11.25 for certain classes of sprays on manifolds have been given in Del Riego and Parker (1995).

CHAPTER 12

SINGULARITIES

A common assumption made in studying Riemannian manifolds is that the spaces under consideration are Cauchy complete or, equivalently, geodesically complete. This assumption seems reasonable since a large number of important Riemannian manifolds are complete.

The situation for Lorentzian manifolds is quite different. A large number of the more important Lorentzian manifolds used as models in general relativity fail to be geodesically complete. Also, the problem of completeness is further complicated by the fact, observed in earlier chapters, that there are a number of inequivalent forms of completeness for Lorentzian manifolds.

In this chapter we will be concerned with establishing theorems which guarantee the nonspacelike geodesic incompleteness of a large class of space–times. These space–times contain at least one nonspacelike geodesic which is both inextendible and incomplete. Such a geodesic has an endpoint $\overline{p}$ in the causal boundary $\partial_c M$ which may be thought of as being outside the space–time but not at infinity. For example, if γ is a future inextendible and future incomplete timelike geodesic which has $\overline{p} \in \partial_c M$ as a future endpoint, then γ corresponds to the path of a "freely falling" test particle which falls to the edge of the universe (at $\overline{p}$) in finite time.

It has been known for a long time in general relativity that a number of important space–times are nonspacelike incomplete. Nonetheless, this incompleteness was thought to be caused by the symmetries of these models. Thus it was felt that nonspacelike completeness was a reasonable assumption for physically realistic space–times. The argument for this assumption was based on physical intuition which was evidently unjustified with hindsight [cf. Tipler, Clarke, and Ellis (1980, Chapter 4)].

If (M, g) is an inextendible space–time which has an inextendible nonspacelike geodesic which is incomplete, then (M, g) is said to have a singularity. The purpose of this chapter is to establish several singularity (i.e., incompleteness) theorems.

Before beginning our study of singularity theory, we pause to explain why, from a mathematical viewpoint, this theory works for all space–times of dimension $n \geq 3$ but *not* for space–times of dimension two. It is simply the fact noted at the beginning of Section 10.3 that *no* null geodesic in a two-dimensional space–time contains any conjugate points. Yet a key argument in proving singularity theorems is showing that certain curvature conditions force every complete nonspacelike geodesic to contain a pair of conjugate points.

Familiarity with the notations and some of the basic properties of Jacobi fields treated in Sections 10.1 and 10.3 will be assumed in this chapter.

12.1 Jacobi Tensors

As we saw in Section 10.3 (Definition 10.61 ff.), Jacobi tensors provide a convenient way of studying conjugate points. Given a timelike geodesic segment $c : [a, b] \to M$, let $N(c(t))$ denote the $(n-1)$-dimensional subspace of $T_{c(t)}M$ consisting of tangent vectors orthogonal to $c'(t)$ as in Definition 10.1. A (1,1) tensor field $A(t)$ on $V^{\perp}(c)$ is a linear map

$$A = A(t) : N(c(t)) \to N(c(t))$$

for each $t \in [a, b]$. Further, a composite endomorphism $RA(t) : N(c(t)) \to N(c(t))$ may be defined by

$$RA(t)(v) = R\left(A(t)(v), c'(t)\right) c'(t).$$

The *adjoint* $A^*(t)$ of $A(t)$ is defined by requiring that

$$g(A(t)(w), v) = g(A^*(t)(v), w)$$

for all $v, w \in N(c(t))$.

A smooth (1,1) tensor field $A(t)$ on $V^{\perp}(c)$ is said to be a *Jacobi tensor field* if

$$A'' + RA = 0$$

and

$$\ker(A(t)) \cap \ker(A'(t)) = \{0\}$$

for all $t \in [a, b]$. If Y is a parallel vector field along c and A is a Jacobi tensor on $V^\perp(c)$, then the vector field $J = A(Y)$ satisfies the differential equation $J'' + R(J, c')c' = 0$ and hence is a Jacobi field. The condition $\ker(A(t)) \cap \ker(A'(t)) = \{0\}$ for all $t \in [a, b]$ guarantees that if Y is a nonzero parallel field along c, then $J = A(Y)$ is a nontrivial Jacobi field. Suppose that A is a Jacobi tensor on $V^\perp(c)$ with $A(a) = 0$. If $A(t_0)(v) = 0$ for some $t_0 \in (a, b]$ and $0 \neq v \in N(c(t_0))$, then letting Y be the unique parallel field along c with $Y(t_0) = v$, we find that $J = A(Y)$ is a Jacobi field with $J(a) = J(t_0) = 0$.

A Jacobi tensor A is said to be a *Lagrange tensor field* if

$$(A')^* A - A^* A' = 0$$

for all $t \in [a, b]$. As in the proof of Lemma 10.67, it may be shown that a Jacobi tensor field A is a Lagrange tensor field if $A(t_0) = 0$ for some $t_0 \in [a, b]$.

Remark 12.1. Let $c : [a, b] \to M$ be a unit speed timelike geodesic, and suppose $E_1, E_2, \ldots, E_n$ is a parallelly propagated orthonormal basis along c with $E_n = c'$. Then $N(c(t))$ is the span of $E_1, E_2, \ldots, E_{n-1}$, and each Jacobi vector field J along c which is everywhere orthogonal to c' may be expressed in terms of $E_1, E_2, \ldots, E_{n-1}$. Thus J may be represented as a column vector with $(n-1)$ components. Using this representation, let $J_i = J_i(t)$ be the column vector corresponding to the Jacobi field J along c which satisfies $J(t_0) = 0$ and $J'(t_0) = E_i(t_0)$. Let

$$A(t) = [J_1(t), J_2(t), \ldots, J_{n-1}(t)]$$

be the $(n-1) \times (n-1)$ matrix with $J_i(t)$ for the ith column. This matrix $A(t)$ is a representation of a Lagrange tensor field along c. Using this same basis $E_1, E_2, \ldots, E_{n-1}$, the adjoint $A^*(t)$ is represented by the transpose of $A(t)$. The space of Jacobi fields which vanish at t_0 and which have derivatives orthogonal to c' at t_0 may be identified with the span of the columns of A. Thus conjugate points of $c(t_0)$ along c are exactly the points where $\det A(t) = 0$.

Hence $\det A(t)$ has isolated zeroes on the interval $[a,b]$. Also, the multiplicity of a conjugate point $t=t_1$ to t_0 along c is just the nullity of $A(t_1): N(c(t_1)) \to N(c(t_1))$.

The fact that $A(t_0)=0$ and $A'(t_0)=E$ is essential to the above discussion. Lagrange tensors along a timelike geodesic $c: J \to (M,g)$ may be constructed which are singular at distinct $t_0, t_1 \in J$, yet c has no conjugate points. For example, let (M,g) be $\mathbb{R}^3$ with the Lorentzian metric $ds^2 = -dx^2 + dy^2 + dz^2$ and let $c(t) = (t,0,0)$. Let $E_1 = \partial/\partial y$ and $E_2 = \partial/\partial z$. Then if A is the Jacobi tensor along c with the matrix representation

$$A(t) = \begin{pmatrix} t & 0 \\ 0 & t-1 \end{pmatrix}$$

with respect to $E_1 \circ c$ and $E_2 \circ c$, we have $A' = E$ and $A^* = A$ so that $(A')^*A - A^*A' = 0$, and A is a Lagrange tensor. Evidently, $A(0)\big(E_1(c(0))\big) = 0$ and $A(1)\big(E_2(c(1))\big) = 0$. But c has no conjugate points since $(\mathbb{R}^3, ds^2)$ is Minkowski 3-space.

We now define the expansion, vorticity, and shear of a Jacobi tensor A along the timelike geodesic $c:[a,b] \to M$. As before, $E = E(t)$ will represent the $(1,1)$ tensor field on $V^\perp(c)$ such that $E(t) = \mathrm{Id} : N(c(t)) \to N(c(t))$ for each t. Note that this definition of B from A parallels the passage in O.D.E. theory from the Jacobi equation to the associated Riccati equation.

Definition 12.2. (*Expansion, Vorticity, and Shear Tensors*) Let A be a Jacobi tensor field along a timelike geodesic, and set $B = A'A^{-1}$ at points where A^{-1} is defined.

(1) The *expansion* θ is defined by

$$\theta = \mathrm{tr}(B).$$

(2) The *vorticity tensor* ω is defined by

$$\omega = \frac{1}{2}(B - B^*).$$

(3) The *shear tensor* σ is defined by

$$\sigma = \frac{1}{2}(B + B^*) - \frac{\theta}{n-1}E.$$

Using an orthonormal basis of parallel fields for $V^\perp(c)$ and matrix algebra, it may be shown that

$$\theta = \operatorname{tr}(A'A^{-1}) = (\det A)^{-1}(\det A)'.$$

Thus if A is a Jacobi tensor field with $A(t_0) = 0$, $A'(t_0) = E$, and $|\theta(t)| \to \infty$ as $t \to t_1$, then $\det A(t_1) = 0$ and $t = t_1$ is conjugate to t_0 along c.

We now calculate the derivative of $B = A'A^{-1}$ using $(A^{-1})' = -A^{-1}A'A^{-1}$. First we have

$$B' = (A'A^{-1})' = A''A^{-1} - A'A^{-1}A'A^{-1} = -R - BB. \tag{12.1}$$

Using $\theta = \operatorname{tr}(B)$ and $B = \omega + \sigma + [\theta/(n-1)]E$, we obtain

$$\begin{aligned}
\theta' &= \operatorname{tr}(B') \\
&= -\operatorname{tr}(R) - \operatorname{tr}(BB) \\
&= -\operatorname{tr}(R) - \operatorname{tr}\left[\left(\omega + \sigma + \frac{\theta}{n-1}E\right)^2\right] \\
&= -\operatorname{tr}(R) - \operatorname{tr}\left(\omega^2 + \sigma^2 + \frac{\theta^2}{(n-1)^2}E\right) \\
&= -\operatorname{tr}(R) - \operatorname{tr}(\omega^2) - \operatorname{tr}(\sigma^2) - \frac{\theta^2}{n-1}
\end{aligned}$$

where we have used $\operatorname{tr}(\omega) = \operatorname{tr}(\sigma) = \operatorname{tr}(\omega\sigma) = 0$. Using the orthonormal basis $E_1, E_2, \ldots, E_n$ along c with $E_n = c'$, we find that

$$\begin{aligned}
\operatorname{tr}(R) &= \sum_{i=1}^{n-1} g(R(E_i, c')c', E_i) \\
&= \sum_{i=1}^{n} g(E_i, E_i)\, g(R(E_i, c')c', E_i) \\
&= \operatorname{Ric}(c', c').
\end{aligned}$$

This yields the *Raychaudhuri equation* for Jacobi tensors along timelike geodesics:

$$\theta' = -\operatorname{Ric}(c', c') - \operatorname{tr}(\omega^2) - \operatorname{tr}(\sigma^2) - \frac{\theta^2}{n-1}.$$

Definition 12.2 implies that the shear tensor σ is self-adjoint for arbitrary Jacobi tensor fields A. Thus if $E_1, E_2, \ldots, E_n$ is an orthonormal basis at $c(t)$ with $E_n = c'(t)$, we may represent σ as a symmetric matrix $[\sigma_{ij}]$ with respect to $E_1, E_2, \ldots, E_{n-1}$. Consequently,

$$\begin{aligned}\operatorname{tr}(\sigma^2) &= \operatorname{tr}\left(\left[\sum_k \sigma_{ik}\sigma_{kj}\right]\right) \\ &= \sum_{i,k} \sigma_{ik}\sigma_{ki} \\ &= \sum_i \sum_k \sigma_{ik}^2 \geq 0.\end{aligned}$$

Thus $\operatorname{tr}(\sigma^2) = 0$ if and only if $\sigma = 0$.

If A is a Lagrange tensor field as well as a Jacobi tensor field, then the tensor $B = A'A^{-1}$ has the following property.

Lemma 12.3. *If A is a Lagrange tensor field, then $B = A'A^{-1}$ is self-adjoint.*

Proof. The equation $A^*A' = A'^*A$ implies that

$$B = A'A^{-1} = A^{*-1}A^{*\prime} = B^*. \quad \square$$

Lemma 12.3 then has the following consequence.

Corollary 12.4. *If A is a Lagrange tensor field, then the vorticity tensor $\omega = (1/2)(B - B^*)$ vanishes along c.*

We thus obtain the *vorticity-free Raychaudhuri equation* for Lagrange tensor fields along timelike geodesics:

$$\theta' = -\operatorname{Ric}(c', c') - \operatorname{tr}(\sigma^2) - \frac{\theta^2}{n-1}. \tag{12.2}$$

We now consider the Raychaudhuri equation for a null geodesic $\beta : [a, b] \to M$. As discussed in Section 10.3, we use the quotient bundle $G(\beta) = N(\beta)/[\beta']$

along β rather than $N(\beta)$. Recall that a smooth $(1,1)$ tensor field $\overline{A} : G(\beta) \to G(\beta)$ is said to be a *Jacobi tensor* along the null geodesic β if

$$\overline{A}'' + \overline{R}\,\overline{A} = 0$$

and

$$\ker(\overline{A}(t)) \cap \ker(\overline{A}'(t)) = \{[\beta'(t)]\}$$

for all $t \in [a,b]$ (cf. Definition 10.61). We proceed in the null case in much the same manner as in the timelike case, remembering that we work modulo β' in $G(\beta)$ and that $\dim G(\beta(t)) = n-2$. Also the adjoint $\overline{A}^*$ of $\overline{A}$ is defined by

$$\overline{g}(\overline{A}w, v) = \overline{g}(\overline{A}^*(v), w)$$

where $\overline{g}$ is the positive definite metric on $G(\beta)$ given by formula (10.31) of Section 10.3.

Definition 12.5. (*Expansion, Vorticity, and Shear Tensors*) Let $\overline{A}$ be a Jacobi tensor along a null geodesic β, and set $\overline{B} = \overline{A}'\,\overline{A}^{-1}$ at points where $\overline{A}^{-1}$ is defined.

(1) The *expansion* $\overline{\theta}$ is defined by

$$\overline{\theta} = \operatorname{tr}(\overline{B}) = (\det \overline{A})^{-1}(\det \overline{A})'.$$

(2) The *vorticity tensor* $\overline{\omega}$ is defined by

$$\overline{\omega} = \frac{1}{2}(\overline{B} - \overline{B}^*).$$

(3) The *shear tensor* $\overline{\sigma}$ is defined by

$$\overline{\sigma} = \frac{1}{2}(\overline{B} + \overline{B}^*) - \frac{\overline{\theta}}{n-2}\overline{E}.$$

Using the same reasoning as in the timelike case we may obtain

(12.3) $$\overline{B}' = -\overline{R} - \overline{B}\,\overline{B}$$

and

$$\overline{\theta}' = -\operatorname{tr}(\overline{R}) - \operatorname{tr}(\overline{\omega}^2) - \operatorname{tr}(\overline{\sigma}^2) - \frac{\overline{\theta}^2}{n-2}.$$

We may calculate $\operatorname{tr}(\overline{R})$ as follows. Let $V(\beta)$ denote the geometric realization for $G(\beta)$ constructed as in (10.28) of Section 10.3, and let $\{Y_1, \ldots, Y_{n-2}\}$ be an orthonormal basis for $V(\beta)$ at every point of β. Extend $\{Y_1, \ldots, Y_{n-2}\}$ to an orthonormal basis $\{Y_1, Y_2, \ldots, Y_n\}$ along β, where Y_n is timelike and $\beta' = (Y_{n-1} + Y_n)/\sqrt{2}$. Then we have

$$\begin{aligned} &g(R(Y_{n-1}, \beta')\beta', Y_{n-1}) - g(R(Y_n, \beta')\beta', Y_n) \\ &= 2^{-1} g(R(Y_{n-1}, Y_n)Y_n, Y_{n-1}) - 2^{-1} g(R(Y_n, Y_{n-1})Y_{n-1}, Y_n) \\ &= 0 \end{aligned}$$

using the basic properties of the curvature tensor. Consequently, we obtain

$$\begin{aligned} \operatorname{tr}(\overline{R}) &= \sum_{i=1}^{n-2} \overline{g}(\overline{R}(\pi(Y_i), \beta')\beta', \pi(Y_i)) \\ &= \sum_{i=1}^{n-2} g(R(Y_i, \beta')\beta', Y_i) \\ &= \sum_{i=1}^{n} g(Y_i, Y_i)\, g(R(Y_i, \beta')\beta', Y_i) \\ &= \operatorname{Ric}(\beta', \beta'). \end{aligned}$$

This yields the *Raychaudhuri equation* for Jacobi tensors along null geodesics:

$$\overline{\theta}' = -\operatorname{Ric}(\beta', \beta') - \operatorname{tr}(\overline{\omega}^2) - \operatorname{tr}(\overline{\sigma}^2) - \frac{\overline{\theta}^2}{n-2}. \tag{12.4}$$

The same reasoning as in the proof of Lemma 12.3 shows we may simplify the above equation when $\overline{A}$ is a Lagrange tensor field (i.e., when $\overline{A}^*\overline{A}' = \overline{A}'^*\overline{A}$).

Lemma 12.6. *If $\overline{A}$ is a Lagrange tensor field, then the vorticity tensor $\overline{\omega}$ vanishes along β'.*

We thus obtain the *vorticity-free Raychaudhuri equation* for Lagrange tensor fields along null geodesics:

$$\overline{\theta}' = -\operatorname{Ric}(\beta', \beta') - \operatorname{tr}(\overline{\sigma}^2) - \frac{\overline{\theta}^2}{n-2}. \tag{12.5}$$

12.2 The Generic and Timelike Convergence Conditions

In this section we show that if (M, g) is a space–time of dimension at least three which satisfies the generic and timelike convergence conditions, then every complete nonspacelike geodesic contains a pair of conjugate points [cf. Hawking and Penrose (1970, p. 539)]. The timelike and null cases are handled separately. Similar treatments of the material in this section may be found in Bölts (1977), Hawking and Ellis (1973, pp. 96–101), and Eschenburg and O'Sullivan (1976).

We may state the definitions of the generic condition and the timelike convergence condition for our purposes in this section as follows (cf. Propositions 2.7 and 2.11).

Definition 12.7. (*Generic Condition*) A timelike geodesic $c : (a, b) \to (M, g)$ is said to satisfy the *generic condition* if there exists some $t_0 \in (a, b)$ such that the curvature endomorphism

$$R(\,\cdot\,, c'(t_0))c'(t_0) \,:\, V^\perp(c(t_0)) \to V^\perp(c(t_0))$$

is not identically zero. A null geodesic $\beta : (a, b) \to (M, g)$ is said to satisfy the *generic condition* if there exists some $t_0 \in (a, b)$ such that the curvature endomorphism

$$\overline{R}(\,\cdot\,, \beta'(t_0))\beta'(t_0) \,:\, G(\beta(t_0)) \to G(\beta(t_0))$$

of the quotient space $G(\beta(t_0))$ is not identically zero. The space–time (M, g) is said to satisfy the *generic condition* if each inextendible nonspacelike geodesic satisfies this condition.

In Section 2.5 we have shown that this formulation of the generic condition is equivalent to the usual definition given in general relativity; namely, a nonspacelike geodesic c with tangent vector W satisfies the generic condition if there is some point of c at which

$$W^c W^d W_{[a} R_{b]cd[e} W_{f]} \neq 0.$$

Definition 12.8. (*Timelike Convergence Condition*) A space–time satisfies the *timelike convergence condition* if $\mathrm{Ric}(v, v) \geq 0$ for all nonspacelike tangent vectors $v \in TM$.

By continuity, the curvature condition of Definition 12.8 is equivalent to the *timelike convergence condition* of Hawking and Ellis (1973, p. 95) that $\mathrm{Ric}(v, v) \geq 0$ for all timelike $v \in TM$. Hawking and Ellis (1973, p. 95) also call the curvature condition $\mathrm{Ric}(w, w) \geq 0$ for all null $w \in TM$ the *null convergence condition.* In Hawking and Ellis (1973, p. 89), a four-dimensional space–time (M, g) with energy-momentum tensor T (cf. Section 2.6) is said to satisfy the *weak energy condition* if $T(v, v) \geq 0$ for all timelike $v \in TM$. If the Einstein equations hold for the four-dimensional space–time (M, g) and T with cosmological constant Λ, then the condition $\mathrm{Ric}(v, v) \geq 0$ for all timelike $v \in TM$ implies that

$$T(v, v) \geq \left(\frac{\mathrm{tr}\, T}{2} - \frac{\Lambda}{8\pi} \right) g(v, v)$$

for all timelike $v \in TM$. Hence in Hawking and Ellis (1973, p. 95), the four-dimensional space–time (M, g) and energy-momentum tensor T are said to satisfy the *strong energy condition* if $T(v, v) \geq (\mathrm{tr}\, T/2) g(v, v)$ for all timelike $v \in TM$. When $\dim M = 4$ and $\Lambda = 0$, this may be seen to be equivalent to the condition $\mathrm{Ric}(v, v) \geq 0$ for all timelike $v \in TM$ (cf. Section 2.6). In Hawking and Penrose (1970, p. 539), the condition $\mathrm{Ric}(v, v) \geq 0$ for all unit timelike vectors $v \in TM$ is called the *energy condition.* In Frankel (1979) and Lee (1975), the term "strong energy condition" is provided the same definition which we have given to "timelike convergence condition" in Definition 12.8. A discussion of the physical interpretation of these curvature conditions in general relativity may be found in Hawking and Ellis (1973, Section 4.3).

As we have just noted above, if (M, g) satisfies the timelike convergence condition or the strong energy condition, then (M, g) satisfies the null convergence condition. In view of several rigidity theorems associated with curvature conditions in Riemannian geometry [cf. Cheeger and Ebin (1975, pp. v and vi)], it is natural to consider the implications of the curvature condition $\mathrm{Ric}(w, w) = 0$ for all null vectors $w \in TM$. Applying linear algebraic arguments to each tangent space, Dajczer and Nomizu (1980a) have obtained the rigidity result that if $\dim M \geq 3$ and $\mathrm{Ric}(w, w) = 0$ for all null vectors $w \in TM$, then (M, g) is Einstein, i.e., $\mathrm{Ric} = \lambda g$ for some constant $\lambda \in \mathbb{R}$. Thus if (M, g) is not Einstein, there are some nonzero null Ricci curvatures. Suppose further that (M, g) is globally hyperbolic with a smooth globally hyperbolic time function $h : M \to \mathbb{R}$ such that for some Cauchy hypersurface $S = h^{-1}(t_0)$, all null Ricci curvatures satisfy $\mathrm{Ric}(g)(w, w) > 0$ if $\pi(w) \in S$. If (M, g) also satisfies the null convergence condition, then M admits a metric g_1 globally conformal to g such that the globally hyperbolic space–time (M, g_1) satisfies the curvature condition $\mathrm{Ric}(g_1)(w, w) > 0$ for all null vectors $w \in TM$ [cf. Beem and Ehrlich (1978, p. 174, Theorem 7.1)].

An essential step in proving that any complete timelike geodesic in a space–time satisfying the generic and timelike convergence conditions contains a pair of conjugate points is the following proposition.

Proposition 12.9. *Let $c : J \to (M, g)$ be an inextendible timelike geodesic satisfying $\mathrm{Ric}(c'(t), c'(t)) \geq 0$ for all $t \in J$. Let A be a Lagrange tensor field along c. Suppose that the expansion $\theta(t) = \mathrm{tr}(A'(t)A^{-1}(t))$ has a negative [respectively, positive] value $\theta_1 = \theta(t_1)$ at $t_1 \in J$. Then $\det A(t) = 0$ for some t in the interval from t_1 to $t_1 - (n-1)/\theta_1$ [respectively, some t in the interval from $t_1 - (n-1)/\theta_1$ to t_1] provided that $t \in J$.*

Proof. Since $\theta = (\det A)'(\det A)^{-1}$, we have $\det A(t_0) = 0$ provided that $|\theta| \to \infty$ as $t \to t_0$. Thus we need only show that $|\theta| \to \infty$ on the above intervals. Put

$$s_1 = \frac{n-1}{\theta_1}.$$

The vorticity-free Raychaudhuri equation (12.2) for timelike geodesics and the condition $\mathrm{Ric}(c', c') \geq 0$ yield the inequality

$$\frac{d\theta}{dt} \leq -\frac{\theta^2}{n-1}.$$

In the case that $\theta_1 < 0$, integrating this inequality from t_1 to $t > t_1$ we obtain

$$\theta(t) \leq \frac{n-1}{t + s_1 - t_1}$$

for $t \in [t_1, t_1 - s_1)$. Hence $|\theta(t)|$ becomes infinite for some $t \in (t_1, t_1 - s_1]$ provided that $c(t)$ is defined. In the case that $\theta_1 > 0$, we obtain for $t \in (t_1 - s_1, t_1]$ that

$$\theta(t) \geq \frac{n-1}{t + s_1 - t_1}.$$

Hence $|\theta(t)|$ again becomes infinite for some $t \in [t_1 - s_1, t_1)$ provided that $c(t)$ is defined. $\square$

We now show that a timelike geodesic in a space–time that satisfies the timelike convergence condition and the generic condition must either be incomplete or else have a pair of conjugate points.

Proposition 12.10. *Let (M, g) be an arbitrary space–time of dimension $n \geq 2$. Suppose that $c : \mathbb{R} \to (M, g)$ is a complete timelike geodesic which satisfies $\mathrm{Ric}(c'(t), c'(t)) \geq 0$ for all $t \in \mathbb{R}$. If $R(\cdot, c'(t_1))c'(t_1) : N(c(t_1)) \to N(c(t_1))$ is not zero for some $t_1 \in \mathbb{R}$, then c has a pair of conjugate points.*

We first prove four lemmas which are needed for the proof of Proposition 12.10 [cf. Bölts (1977, pp. 30–37)].

Lemma 12.11. *Let $c : [a, b] \to (M, g)$ be a timelike geodesic without conjugate points. Then there is a unique $(1, 1)$ tensor field A on $V^{\perp}(c)$ which satisfies the differential equation $A'' + RA = 0$ with given boundary conditions $A(a)$ and $A(b)$.*

Proof. Let S be the vector space of $(1, 1)$ tensor fields A on $V^{\perp}(c)$ with $A'' + RA = 0$, and let $L(N(c(t)))$ denote the set of linear endomorphisms of $N(c(t))$. Define a linear transformation $\phi : S \to L(N(c(a))) \times L(N(c(b)))$ by

$$\phi(A) = (A(a), A(b)).$$

Since $\dim S = \dim\big(L(N(c(a)))\big) + \dim\big(L(N(c(b)))\big) = 2(n-1)^2$, it is only necessary to show that ϕ is injective in order to prove that ϕ is an isomorphism and establish the existence of a unique solution A. Assume that $\phi(A) = (A(a), A(b)) = (0, 0)$. If $Y(t)$ is any parallel vector field along c, then $J(t) = A(t)Y(t)$ is a Jacobi field with $J(a) = J(b) = 0$. Thus $J = 0$. However, since $Y(t)$ was an arbitrary parallel field, this implies that $A(t) = 0$, which shows ϕ is injective and establishes the lemma. $\square$

Now let $c : [t_1, \infty) \to (M, g)$ be a timelike geodesic without conjugate points and fix $s \in (t_1, \infty)$. Then by Lemma 12.11, there exists a unique $(1, 1)$ tensor field on $V^\perp(c)$, which we will denote by D_s, satisfying the differential equation $D_s{}'' + RD_s = 0$ with initial conditions $D_s(t_1) = E$ and $D_s(s) = 0$. As $D_s(t_1) = E$, we have $\ker(D_s(t_1)) \cap \ker(D_s{}'(t_1)) = \{0\}$. Thus D_s is a Jacobi tensor field (cf. Lemma 10.62). Also, since $D_s(s) = 0$, it follows that D_s is a Lagrange tensor field. It will also be shown during the course of the proof of Lemma 12.12 that if A is the Lagrange tensor field on $V^\perp(c)$ with $A(t_1) = 0$ and $A'(t_1) = E$, then $D_s{}'(s) = -(A^*)^{-1}(s)$.

Lemma 12.12. *Let $c : [t_1, \infty) \to M$ be a timelike geodesic without conjugate points. Let A be the unique Lagrange tensor on $V^\perp(c)$ with $A(t_1) = 0$ and $A'(t_1) = E$. Then for each $s \in (t_1, \infty)$ the Lagrange tensor D_s on $V^\perp(c)$ with $D_s(t_1) = E$ and $D_s(s) = 0$ satisfies the equation*

$$D_s(t) = A(t) \int_t^s (A^*A)^{-1}(\tau)\, d\tau$$

for all $t \in (t_1, s]$. Thus $D_s(t)$ is nonsingular for $t \in (t_1, s)$.

Proof. Set $X(t) = A(t) \int_t^s (A^*A)^{-1}(\tau)\, d\tau$. It suffices to show that $X'' + RX = 0$, $X(s) = D_s(s) = 0$, and $X'(s) = D_s{}'(s)$.

We first check that $X'' + RX = 0$. Differentiating, we obtain

$$\begin{aligned} X'(t) &= A'(t) \int_t^s (A^*A)^{-1}(\tau)\, d\tau - A(t)(A^*A)^{-1}(t) \\ &= A'(t) \int_t^s (A^*A)^{-1}(\tau)\, d\tau - (A^*)^{-1}(t). \end{aligned}$$

Hence

$$\begin{aligned} X''(t) &= A''(t) \int_t^s (A^*A)^{-1}(\tau)\, d\tau - A'(t)(A^*A)^{-1}(t) \\ &\quad - \left((A^*)^{-1}\right)'(t) \\ &= A''(t) \int_t^s (A^*A)^{-1}(\tau)\, d\tau - A'(t)A^{-1}(t)(A^*)^{-1}(t) \\ &\quad + (A^*)^{-1}(A^*)'(A^*)^{-1}(t). \end{aligned}$$

But since A is a nonsingular Lagrange tensor, $(A^*)' = A^*A'A^{-1}$, so that $(A^*)^{-1}(A^*)'(A^*)^{-1} = A'A^{-1}(A^*)^{-1}$, and we obtain

$$X''(t) = A''(t) \int_t^s (A^*A)^{-1}(\tau)\, d\tau.$$

We then have

$$X''(t) + R(t)X(t) = [A''(t) + R(t)A(t)] \int_t^s (A^*A)^{-1}(\tau)\, d\tau = 0$$

since $A''(t) + R(t)A(t) = 0$. Thus X satisfies the Jacobi differential equation.

Setting $t = s$, we obtain

$$X(s) = A(s) \int_s^s (A^*A)^{-1}(\tau)\, d\tau = 0$$

and

$$X'(s) = A'(s) \int_s^s (A^*A)^{-1}(\tau)\, d\tau - A(s)(A^*A)^{-1}(s) = -(A^*)^{-1}(s).$$

Thus it remains to check that $D_s{}'(s) = -(A^*)^{-1}(s)$. But using $R^* = R$, we obtain

$$\begin{aligned}[(A^*)'D_s - A^*D_s{}']' &= (A^*)''D_s + (A^*)'D_s{}' - (A^*)'D_s{}' - A^*D_s{}'' \\ &= (A^*)''D_s - A^*D_s{}'' \\ &= -A^*R^*D_s + A^*RD_s = 0.\end{aligned}$$

Thus $(A^*)'D_s - A^*D_s{}'$ is parallel along c. At $t = t_1$, the initial conditions $A(t_1) = 0$ and $A'(t_1) = E$ for A imply that $A^*(t_1) = 0$ and $(A^*)'(t_1) = (A')^*(t_1) = E$. Hence as $D_s(t_1) = E$, we obtain

$$((A^*)'D_s - A^*D_s{}')(t_1) = E.$$

Hence $((A^*)'D_s - A^*D_s{}')(t) = E$ for all t. Setting $t = s$, we have

$$E = ((A^*)'D_s - A^*D_s{}')(s) = -(A^*D_s{}')(s)$$

which implies that $D_s{}'(s) = -(A^*)^{-1}(s) = X'(s)$. Therefore since D_s and X both satisfy $A'' + RA = 0$ and have the same values and first derivatives at $t = s$, the tensors must agree for all t.

Finally, the nonsingularity of $D_s(t)$ for $t \in (t_1, s)$ follows from the formula

$$D_s(t) = A(t) \int_t^s (A^* A)^{-1}(\tau)\, d\tau$$

since $(A^*A)^{-1}(t)$ is a positive definite, self-adjoint tensor field for all $t > t_1$. □

Note that while the integral representation of the Lagrange tensor D_s along c satisfying $D_s(t_1) = E$ and $D_s(s) = 0$ given in Lemma 12.12 was proven only for $t \in (t_1, s]$, if c is defined for all $t \in \mathbb{R}$ and has no conjugate points, then $D_s(t)$ is defined for all $t \in \mathbb{R}$.

We now show that if $c : [a, \infty) \to (M, g)$ is a timelike geodesic without conjugate points, then the above tensor fields D_s converge to a Lagrange tensor field D as $s \to \infty$. This construction parallels the construction of stable Jacobi fields in certain classes of complete Riemannian manifolds without conjugate points [cf. Eschenburg and O'Sullivan (1976, pp. 227 ff.), Green (1958), E. Hopf (1948, p. 48)].

Lemma 12.13. *Let $c : [a, \infty) \to (M, g)$ be a timelike geodesic without conjugate points. For $t_1 > a$ and $s \in [a, \infty) - \{t_1\}$, let D_s be the Lagrange tensor field along c determined by $D_s(t_1) = E$ and $D_s(s) = 0$. Then $D(t) = \lim_{s \to \infty} D_s(t)$ is a Lagrange tensor field. Furthermore, $D(t)$ is nonsingular for all t with $t_1 < t < \infty$.*

Proof. We first show that $D_s'(t_1)$ has a self-adjoint limit as $s \to \infty$. Since D_s is a Lagrange tensor, $(D_s'^* D_s)(t_1) = (D_s^* D_s')(t_1)$. Using $D_s(t_1) = E$, we obtain $D_s'^*(t_1) = D_s'(t_1)$. Thus the limit of $D_s'(t_1)$ must be a self-adjoint linear map which we will denote by $D'(t_1) : N(c(t_1)) \to N(c(t_1))$ if it exists. Consequently, we need only show that for each $y \in N(c(t_1))$, the value of $g(D_s'(t_1)y, y)$ converges to some value $g(D'(t_1)y, y)$.

We will show that the function $s \to g(D_s'(t_1)y, y)$ is monotone increasing for all s with $t_1 < s < \infty$ and is bounded from above by $g(D_a'(t_1)y, y)$ to establish the existence of this limit. To this end, assume that $t_1 < r < s$. Then by Lemma 12.12 we have

$$D_s'(t) = A'(t) \int_t^s (A^* A)^{-1}(\tau)\, d\tau - (A^*)^{-1}(t).$$

Thus for $t \in (t_1, s)$ we obtain

$$g(D_s{}'(t)Y(t), Y(t)) = g\left(\left[A'(t)\int_t^s (A^*A)^{-1}(\tau)\,d\tau\right](Y(t)), Y(t)\right) \\ - g\left((A^*)^{-1}(t)Y(t), Y(t)\right)$$

where A is the Lagrange tensor field along c satisfying $A(t_1) = 0$, $A'(t_1) = E$, and $Y(t)$ is the parallel vector field along c with $Y(t_1) = y$. Thus for t with $t_1 < t < r$, it follows that

$$g(D_s{}'(t)Y(t), Y(t)) - g(D_r{}'(t)Y(t), Y(t))$$

is given by

$$g\left(\left[A'(t)\int_r^s (A^*A)^{-1}(\tau)\,d\tau\right](Y(t)), Y(t)\right).$$

Letting $t \to t_1^+$ and using $Y(t_1) = y$ and $A'(t_1) = E$, we then have

$$g(D_s{}'(t_1)y, y) - g(D_r{}'(t_1)y, y) = g\left(\left[\int_r^s (A^*A)^{-1}(\tau)\,d\tau\right](Y(t_1)), Y(t_1)\right).$$

Since Y is parallel along c, it may be checked by choosing an orthonormal basis of parallel fields for $V^\perp(c)$ that

$$g\left(\left[\int_r^s (A^*A)^{-1}(\tau)\,d\tau\right](Y(t_1)), Y(t_1)\right) \\ = \int_r^s g\left((A^*A)^{-1}(\tau)(Y(\tau)), Y(\tau)\right)\,d\tau.$$

Since $(A^*A)^{-1} = A^{-1}A^{*-1}$, this may be written as

$$\int_r^s g\left((A^*)^{-1}(\tau)Y(\tau), (A^*)^{-1}(\tau)Y(\tau)\right)\,d\tau$$

which must be positive because $(A^*)^{-1}(\tau)Y(\tau)$ is a spacelike vector in $N(c(\tau))$ for each $\tau \in [r, s]$. Thus

$$g(D_s{}'(t_1)y, y) - g(D_r{}'(t_1)y, y) > 0,$$

and the map $s \to g(D_s{}'(t_1)y, y)$ is monotone for all $s > t_1$ as required.

We now show that $g(D_s{}'(t_1)y, y) < g(D_a{}'(t_1)y, y)$ for all $s > t_1$ and any $y \in N(c(t_1))$. Again let Y be the unique parallel field along c with $Y(t_1) = y$. Let J be the piecewise smooth Jacobi field along $c \mid [a, s]$ given by

$$J(t) = \begin{cases} D_a(t)Y(t) & \text{for } a \leq t < t_1, \\ D_s(t)Y(t) & \text{for } t_1 \leq t \leq s. \end{cases}$$

Also let $J_a = J \mid [a, t_1]$ and $J_s = J \mid [t_1, s]$. Then $J(a) = J(s) = 0$, and J is well-defined at $t = t_1$ since $D_a(t_1) = D_s(t_1) = E$. Using the index form I for $c \mid [a, s]$ given in Definition 10.4 of Section 10.1, we obtain

$$\begin{aligned} I(J, J) &= I(J, J)_a^{t_1} + I(J, J)_{t_1}^s \\ &= -g(J_a{}'(t_1), J_a(t_1)) + g(J_s{}'(t_1), J_s(t_1)) \\ &= -g(D_a{}'(t_1)Y(t_1), Y(t_1)) + g(D_s{}'(t_1)Y(t_1), Y(t_1)) \\ &= -g(D_a{}'(t_1)y, y) + g(D_s{}'(t_1)y, y) \end{aligned}$$

where we have used formula (10.2) of Definition 10.4 and $D_a(t_1) = D_s(t_1) = E$. Since $J(a) = J(s) = 0$ and c has no conjugate points in $[a, \infty)$, we have $I(J, J) < 0$ by Theorem 10.22. Thus

$$g(D_s{}'(t_1)y, y) < g(D_a{}'(t_1)y, y)$$

for all $s > t_1$; we conclude that the self-adjoint tensor $D'(t_1) = \lim\limits_{s \to \infty} D_s{}'(t_1)$ exists.

Now we define $D(t)$ by setting $D(t)$ equal to the unique Jacobi tensor along c which satisfies $D(t_1) = E$ and $D'(t_1) = \lim_{s \to \infty} D_s{}'(t_1)$. Since $D(t)$ and $D_s(t)$ both satisfy the differential equation $A'' + RA = 0$ and the initial conditions of D_s approach the initial conditions of D as $s \to \infty$, it follows that $D(t) = \lim_{s \to \infty} D_s(t)$ and $D'(t) = \lim_{s \to \infty} D_s{}'(t)$ for all $t \in [a, \infty)$. This implies that the limit $D(t)$ of the Lagrange tensors $D_s(t)$ must also be a Lagrange tensor.

The last statement of the lemma now follows using the representation

$$D(t) = A(t) \int_t^\infty (A^* A)^{-1}(\tau)\, d\tau$$

and the fact that $(A^* A)^{-1}(t)$ is a positive definite self-adjoint tensor field for all $t > t_1$. $\square$

Now divide the Lagrange tensors with $A(t_1) = E$ along a complete timelike geodesic $c: (-\infty, +\infty) \to (M, g)$ with $\text{Ric}(c', c') \geq 0$ and $R(\cdot, c'(t_1))c'(t_1) \neq 0$ for some $t_1 \in \mathbb{R}$ into two classes L_+ and L_- as follows [cf. Bölts (1977, p. 36), Hawking and Ellis (1973, p. 98)]. Put

$$L_+ = \{A : A \text{ is a Lagrange tensor with } A(t_1) = E \text{ and} \\ \theta(t_1) = \text{tr}(A'(t_1)) \geq 0\}$$

and

$$L_- = \{A : A \text{ is a Lagrange tensor with } A(t_1) = E \text{ and} \\ \theta(t_1) = \text{tr}(A'(t_1)) \leq 0\}$$

Lemma 12.14. *Let $c : \mathbb{R} \to (M, g)$ be a complete timelike geodesic such that $\text{Ric}(c', c') \geq 0$ and $R(\cdot, c'(t_1))c'(t_1) \neq 0$ for some $t_1 \in \mathbb{R}$. Then each $A \in L_-$ satisfies $\det A(t) = 0$ for some $t > t_1$, and each $A \in L_+$ satisfies $\det A(t) = 0$ for some $t < t_1$.*

Proof. If $A \in L_-$, then $\theta(t_1) = \text{tr}(A'(t_1)A^{-1}(t_1)) = \text{tr}(A'(t_1)) \leq 0$. Using the vorticity-free Raychaudhuri equation (12.2) for timelike geodesics with $\text{Ric}(c', c') \geq 0$ and $\text{tr}(\sigma^2) \geq 0$, we find $\theta'(t) \leq 0$ for all t. Thus $\theta(t) \leq 0$ for all $t \geq t_1$. If $\theta(t_0) < 0$ for some $t_0 > t_1$, then the result for A follows from Proposition 12.9. Assume therefore that $\theta(t) = 0$ for $t \geq t_1$. This implies $\theta'(t) = 0$ for $t \geq t_1$ which yields $\text{tr}(\sigma^2) = 0$. Hence $\sigma = 0$ for $t \geq t_1$ since σ is self-adjoint. Using $\theta = 0$ and the self-adjointness of B, we thus have $B = \sigma = 0$ which by equation (12.1) implies that $R = -B^2 - B' = 0$ for $t \geq t_1$, in contradiction to $R(t_1) \neq 0$.

If $A \in L_+$, the proof is similar. □

We now come to the

Proof of Proposition 12.10. Let $c : \mathbb{R} \to (M, g)$ be a complete timelike geodesic with $\text{Ric}(c'(t), c'(t)) \geq 0$ for all $t \in \mathbb{R}$ and with $R(\cdot, c'(t_1))c'(t_1) \neq 0$ for some $t_1 \in \mathbb{R}$. Suppose that c has no conjugate points. Then let $D = \lim_{s \to \infty} D_s$ be the Lagrange tensor field on $V^\perp(c)$ with $D(t_1) = E$ constructed in Lemma 12.13. Since $c\,|\,[t_1,\infty)$ has no conjugate points, $D(t)$ is nonsingular for all $t \geq t_1$. Thus $D \notin L_-$ by Lemma 12.14. Hence $D \in L_+$ and, moreover, $\operatorname{tr} D'(t_1) > 0$ as $D \notin L_-$. Since $D'(t_1) = \lim_{s\to\infty} D_s'(t_1)$, there exists an $s > t_1$ such that $\operatorname{tr}(D_s'(t_1)) > 0$. Hence by Lemma 12.14, there exist a $t_2 < t_1$ and a nonzero tangent vector $v \in N(c(t_2))$ such that $D_s(t_2)(v) = 0$. Recall also from the proof of Lemma 12.12 that $D_s(s) = 0$ but $D_s'(s) = -(A^*)^{-1}(s)$ is nonsingular. Therefore, if we let $Y \in V^\perp(c)$ be the unique parallel field along c with $Y(t_2) = v$, then $J = D_s(Y)$ is a nontrivial Jacobi field along c with $Y(t_2) = Y(s) = 0$, in contradiction. □

Corollary 12.15. *Let (M,g) be a space–time of dimension $n \geq 2$ which satisfies the timelike convergence condition and the generic condition. Then each timelike geodesic of (M,g) is either incomplete or else has a pair of conjugate points.*

We now consider the existence of conjugate points on null geodesics. The methods and results for null geodesics are much the same as for timelike geodesics except that it is now necessary to assume that $\dim M \geq 3$ since $\dim G(\beta) = \dim M - 2$ and null geodesics in two-dimensional space–times are free of conjugate points. First, using the vorticity-free Raychaudhuri equation (12.5) for null geodesics, the following analogue of Proposition 12.9 may be established with the same type of reasoning as in the timelike case.

Proposition 12.16. *Let (M,g) be an arbitrary space–time of dimension $n \geq 3$. Suppose that $\beta : J \to (M,g)$ is an inextendible null geodesic satisfying $\operatorname{Ric}(\beta'(t),\beta'(t)) \geq 0$ for all $t \in J$. Let $\overline{A}$ be a Lagrange tensor field along β such that the expansion $\overline{\theta}(t) = \operatorname{tr}(\overline{A}'(t)\,\overline{A}^{-1}(t)) = [\det \overline{A}(t)]^{-1}[\det \overline{A}(t)]'$ has the negative [respectively, positive] value $\overline{\theta}_1 = \overline{\theta}(t_1)$ at $t_1 \in J$. Then $\det \overline{A}(t) = 0$ for some t in the interval from t_1 to $t_1 - (n-2)/\overline{\theta}_1$ [respectively, some t in the interval from $t_1 - (n-2)/\overline{\theta}_1$ to t_1] provided that $t \in J$.*

The null analogue of Proposition 12.10 may also be established for all space–times of dimension $n \geq 3$ using $\overline{A}$, $\overline{B}$, $\overline{\theta}$, $\overline{\sigma}$, etc. in place of the corresponding A, B, θ, σ, etc. used in the proof for timelike geodesics.

Proposition 12.17. *Let $\beta : \mathbb{R} \to (M, g)$ be a complete null geodesic with* $\mathrm{Ric}(\beta'(t), \beta'(t)) \geq 0$ *for all* $t \in \mathbb{R}$*. If* $\dim M \geq 3$ *and if* $\overline{R}(\cdot, \beta'(t))\beta'(t) : G(\beta(t)) \to G(\beta(t))$ *is nonzero for some* $t_1 \in \mathbb{R}$*, then* β *has a pair of conjugate points.*

Combining this result with Corollary 12.15, we obtain the following theorem.

Theorem 12.18. *Let (M, g) be a space–time of dimension $n \geq 3$ which satisfies the timelike convergence condition and the generic condition. Then each nonspacelike geodesic in (M, g) is either incomplete or else has a pair of conjugate points. Thus every nonspacelike geodesic in (M, g) without conjugate points is incomplete.*

The material presented in this section may also be treated within the framework of conjugate points and oscillation theory in ordinary differential equations [cf. Tipler (1977d, 1978), Chicone and Ehrlich (1980), Ehrlich and Kim (1994)]. In this approach, the Raychaudhuri equation is transformed by a change of variables to the differential equation

$$x''(t) + F(t)x(t) = 0$$

with

$$F(t) = \tfrac{1}{m}[\mathrm{Ric}(\gamma'(t), \gamma'(t)) + 2\sigma^2(t)]$$

where $m = n - 1$ if γ is timelike and $m = n - 2$ if γ is null.

12.3 Focal Points

The concept of a conjugate point along a geodesic can be generalized to the notion of a focal point of a submanifold. Let H be a nondegenerate submanifold of the space–time (M, g). At each $p \in H$ the tangent space T_pH may be naturally identified with the vectors of T_pM which are tangent to H at p. The normal space $T_p^{\perp}H$ consists of all vectors orthogonal to H at p. Since H is nondegenerate, $T_p^{\perp}H \cap T_pH = \{0_p\}$ for all $p \in H$. We will denote the exponential map restricted to the normal bundle $T^{\perp}H$ by $\exp^{\perp}$. Then the vector $X \in T_p^{\perp}H$ is said to be a *focal point* of H if $(\exp^{\perp})_*$ is singular at X.

The corresponding point $\exp^{\perp}(X)$ of M is said to be a *focal point of H along the geodesic segment* $\exp^{\perp}(tX)$. When H is a single point, then $T_p^{\perp}H = T_pM$, and a focal point is just an ordinary conjugate point.

Focal points may also be defined using Jacobi fields and the second fundamental form [cf. Bishop and Crittenden (1964, p. 225)]. This approach will be used in this section following the treatment given in Bölts (1977). Jacobi fields are used to measure the separation (or deviation) of nearby geodesics. For example, when a point q is conjugate to p along a geodesic c, geodesics which start at p with initial tangent close to c' at p will tend to focus at q up to second order. They need not actually pass through q but must pass close to q. In studying submanifolds one may take a congruence of geodesics orthogonal to the submanifold and use Jacobi fields to measure the separation of geodesics in this congruence. If p is a focal point along a geodesic c which is orthogonal to the submanifold H, then some geodesics close to c and orthogonal to H tend to focus at p. This is illustrated for the Euclidean plane with the usual positive definite metric in Figure 12.1 and for Lorentzian manifolds in Figure 12.2.

In Section 3.5 we defined the second fundamental form $S_n : T_pH \times T_pH \to \mathbb{R}$ in the direction n, the second fundamental form $S : T_p^{\perp}H \times T_pH \times T_pH \to \mathbb{R}$ and the second fundamental form operator $L_n : T_pH \to T_pH$ (cf. Definition 3.48). Recall that $S(n,x,y) = S_n(x,y) = S_n(y,x)$ and $g(L_n(x),y) = S_n(x,y) = g(\nabla_X Y, n)$ for $n \in T_p^{\perp}H$ and $x, y \in T_pH$, where X and Y are local vector field extensions of x, y.

In this section we will primarily be concerned with the operator $L_n : T_pH \to T_pH$. Note that a vector field η which is orthogonal to H at all points of H defines a $(1,1)$ tensor field L_η on H. We will first consider spacelike hypersurfaces. If the timelike normal vector field η on H satisfies $g(\eta,\eta) = -1$, then we may calculate L_η as follows.

Lemma 12.19. *Let H be a spacelike hypersurface with unit timelike normal field η. If x is a vector tangent to H, then $L_\eta(x) = -\nabla_x \eta$.*

Proof. Since $g(\eta,\eta) = -1$, we have $0 = x(g(\eta,\eta)) = 2g(\nabla_x\eta,\eta)$ which shows that $\nabla_x\eta$ is tangent to H. Now if Y is any vector field tangent to

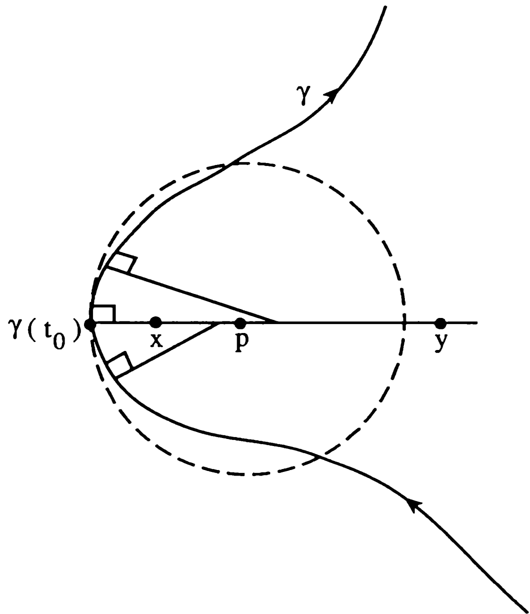


FIGURE 12.1. A curve in the Euclidean plane has its centers of curvature as its focal points. The osculating circle to the curve γ at $t = t_0$ is shown. The segment from $\gamma(t_0)$ to the center p of the osculating circle contains x in its interior. The point y lies beyond p on the ray from $\gamma(t_0)$ through p. For some interval $t_0 - \epsilon_1 < t < t_0 + \epsilon_1$, the closest point of γ to x is $\gamma(t_0)$. On the other hand, for some interval $t_0 - \epsilon_2 < t < t_0 + \epsilon_2$, the farthest point of $\gamma(t)$ from y is $\gamma(t_0)$. Furthermore, the straight lines which are orthogonal to γ near $\gamma(t_0)$ tend to focus at p.

H, then $g(\eta, Y) = 0$. Thus $0 = x(g(\eta, Y)) = g(\nabla_x \eta, Y) + g(\eta, \nabla_x Y)$ so that $g(\nabla_x Y, \eta) = g(-\nabla_x \eta, Y)$. Consequently, $g(L_\eta(x), Y) = g(\nabla_x Y, \eta) = g(-\nabla_x \eta, Y)$. Since Y was arbitrary, the result follows. □

Given a unit normal field η for the spacelike hypersurface H, the collection of unit speed timelike geodesics orthogonal to H with initial direction $\eta(q)$

at $q \in H$ determines a congruence of timelike geodesics. Let c be a timelike geodesic of this congruence intersecting H at q, and let J denote the variation vector field along c of a one-parameter subfamily of the congruence. Then J is a Jacobi field which measures the rate of separation of geodesics in the one-parameter subfamily from c. Since the geodesics of the congruence are all orthogonal to H, it may be shown using Lemma 12.19 that J satisfies the initial condition

$$J'(q) = -L_{\eta(q)} J.$$

This suggests the following definition of a focal point to a spacelike hypersurface in terms of Jacobi fields.

Definition 12.20. (*Focal Point on a Timelike Geodesic*) Let c be a timelike geodesic which is orthogonal to the spacelike hypersurface H at q. A point p on c is said to be a *focal point of H along c* if there is a nontrivial Jacobi field J along c such that J is orthogonal to c', vanishes at p, and satisfies $J' = -L_\eta J$ at q.

Suppose that A is a Jacobi tensor along the timelike geodesic c which satisfies $A = E$ and $A' = -L_\eta A = -L_\eta$ at the point q, where c intersects the spacelike hypersurface H. Then prior to the first focal point, every Jacobi field J orthogonal to c which satisfies $J' = -L_\eta J$ at q may be expressed as $J = AY$ where $Y = Y(t)$ is a parallel vector field along c which is orthogonal to c. Since there are $n - 1$ linearly independent parallel vector fields orthogonal to c, there is an $(n-1)$-dimensional vector space of Jacobi fields along c which satisfy $J' = -L_\eta J$ at q. We now show that such a Jacobi tensor A satisfying $A = E$ and $A' = -L_\eta$ at q is in fact a Lagrange tensor field.

Lemma 12.21. *Suppose that A is a Jacobi tensor field along the timelike geodesic c. Let c be orthogonal to H at t_1, and let L_η be the second fundamental form operator on H. If $A(t_1) = E$ and $A'(t_1) = -L_\eta A(t_1)$, then A is a Lagrange tensor field.*

Proof. The second fundamental form S_η at $\eta \in T^\perp H$ is symmetric, which implies that L_η is self-adjoint at q since $g(L_\eta(x), y) = S_\eta(x, y) = S_\eta(y, x) =$

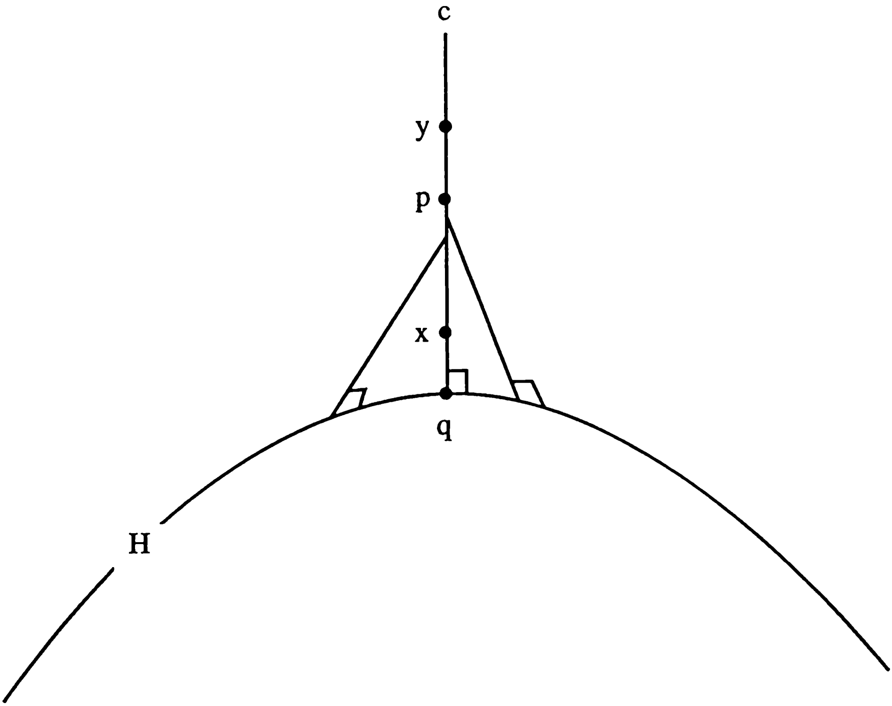


FIGURE 12.2. A spacelike submanifold H in a Lorentzian manifold (M, g) is shown. Here p is a focal point of H along the geodesic c. The geodesic segment from p to q contains x in its interior, and y lies beyond p on the geodesic c. All nonspacelike curves "close" to $c[q, x]$ which join a point of H near q to x have length less than or equal to $c[q, x]$. On the other hand, the farthest point of H near q to y is not q. There are points close to q on H which may be joined to y by timelike curves longer than $c[q, y]$. Furthermore, there is at least one curve γ on H through q such that a family of geodesics orthogonal to H with respect to the given Lorentzian metric and starting on γ near q tend to focus at p up to second order.

$g(L_\eta(y), x)$. Thus at t_1 we have that $A'(t_1) = -L_\eta A(t_1) = -L_\eta$ is self-adjoint. Hence $A^{*\prime}(t_1) = A'(t_1)$. Using $A(t_1) = A^*(t_1) = E$, it follows that $A'(t_1)^* A(t_1) = A^{*\prime}(t_1)A(t_1) = A^*(t_1)A'(t_1)$. Thus A is a Lagrange tensor field as required. □

The tensor $B = A'A^{-1}$, expansion θ, and shear σ of Lagrange tensors A satisfying the conditions of Lemma 12.21 may be defined as in Section 12.1, Definition 12.2. As before, the expansion θ of the Lagrange tensor A along c satisfies the vorticity-free Raychaudhuri equation (12.2) for timelike geodesics. We now establish the following analogue of Proposition 12.9 for spacelike hypersurfaces.

Proposition 12.22. *Let (M, g) be an arbitrary space–time of dimension $n \geq 2$. Suppose that $c : J \to (M, g)$ is an inextendible timelike geodesic which satisfies* $\operatorname{Ric}(c'(t), c'(t)) \geq 0$ *for all $t \in J$ and is orthogonal to the spacelike hypersurface H at $q = c(t_1)$. If $-\operatorname{tr}(L_\eta)$ has the negative [respectively, positive] value θ_1 at q, then there is a focal point $c(t)$ to H for t in the interval from t_1 to $t_1 - (n-1)/\theta_1$, [respectively, in the interval from $t_1 - (n-1)/\theta_1$ to t_1] provided that $t \in J$.*

Proof. By Lemma 12.21, the Jacobi tensor field A along c with initial conditions $A(t_1) = E$ and $A'(t_1) = -L_{\eta(q)}$ is a Lagrange tensor field. Also, $\theta_1 = \theta(t_1) = -\operatorname{tr}(L)$. Hence by Proposition 12.9 the tensor A is singular on the interval from t_1 to $t_1 - (n-1)/\theta_1$. Thus the result follows from the remarks following Definition 12.20. □

For the purpose of studying focal points to submanifolds, it will be helpful to have the second variation formula for the arc length functional at hand. We thus give a derivation of the first and second variation formulas for completeness. Consider a piecewise smooth variation $\alpha : [a, b] \times (-\epsilon, \epsilon) \to (M, g)$ of a piecewise smooth timelike curve $c : [a, b] \to (M, g)$. Hence $\alpha(t, 0) = c(t)$ for all $t \in [a, b]$, and there is a finite partition $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ such that $\alpha \,|\, [t_{i-1}, t_i] \times (-\epsilon, \epsilon)$ is a smooth variation of $c \,|\, [t_{i-1}, t_i]$ for each $i = 1, 2, \ldots, k$ (cf. Definition 10.6). We will also assume that the neighboring curves $\alpha_s = \alpha(\,\cdot\,, s) : [a, b] \to (M, g)$ are timelike for all s with $-\epsilon < s < \epsilon$

(cf. Lemma 10.7). As in Section 10.1, for t with $t_{i-1} \leq t < t_i$, we define the variation vector field V of α along c by

$$V(t) = (\alpha\,|\,[t_{i-1}, t_i] \times (-\epsilon, \epsilon))_* \left.\frac{\partial}{\partial s}\right|_{(t,0)} = \left.\frac{d}{ds}(\alpha(t,s))\right|_{s=0}$$

Also, set

$$\Delta_{t_i}(Y') = Y'(t_i^+) - Y'(t_i^-) \qquad \text{for } i = 1, 2, \ldots, k-1,$$

$$\Delta_{t_k}(Y') = -Y'(t_k^-), \qquad \text{and} \qquad \Delta_{t_0}(Y') = Y'(t_0^+)$$

for any piecewise smooth vector field $Y(t)$ along c which is smooth on each subinterval (t_{i-1}, t_i).

As in Section 10.1, we will use $L(s) = L(\alpha_s)$ to denote the length of the curve $t \mapsto \alpha(t,s)$. The first variation formula for the arc length functional may then be derived as usual.

Proposition 12.23. *Let $c : [a,b] \to (M,g)$ be a unit speed timelike curve which is piecewise smooth. If $\alpha : [a,b] \times (-\epsilon, \epsilon) \to (M,g)$ is a variation of c through timelike curves, then*

$$L'(0) = \int_a^b g(V, c'')|_{(t,0)}\, dt + \sum_{i=0}^{k} g(V(t_i), \Delta_{t_i}(c')).$$

Proof. If $L_i : (-\epsilon, \epsilon) \to \mathbb{R}$ denotes the arc length function of $\alpha\,|\,[t_{i-1}, t_i] \times \{s\}$, then $L(s) = \sum_{i=1}^{k} L_i(s)$ and

$$L_i(s) = \int_{t_{i-1}}^{t_i} \sqrt{-g\left(\alpha_* \frac{\partial}{\partial t}, \alpha_* \frac{\partial}{\partial t}\right)}\, dt.$$

Thus

$$\text{(12.6)} \quad \frac{dL_i}{ds} = \int_{t_{i-1}}^{t_i} \frac{1}{2}\left[-g\left(\alpha_* \frac{\partial}{\partial t}, \alpha_* \frac{\partial}{\partial t}\right)\right]^{-1/2} \left[-2g\left(\nabla_{\partial/\partial t}\left(\alpha_* \frac{\partial}{\partial s}\right), \alpha_* \frac{\partial}{\partial t}\right)\right] dt.$$

On the other hand, since

$$\frac{d}{dt} g\left(\alpha_* \frac{\partial}{\partial s}, \alpha_* \frac{\partial}{\partial t}\right) = g\left(\nabla_{\partial/\partial t}\alpha_* \frac{\partial}{\partial s}, \alpha_* \frac{\partial}{\partial t}\right) + g\left(\alpha_* \frac{\partial}{\partial s}, \nabla_{\partial/\partial t}\alpha_* \frac{\partial}{\partial t}\right)$$

and

$$g\left(\alpha_* \frac{\partial}{\partial t}, \alpha_* \frac{\partial}{\partial t}\right)\Big|_{(t,0)} = -1,$$

it follows that

$$\frac{dL_i}{ds}\Big|_{s=0} = \int_{t_{i-1}}^{t_i} g\left(\alpha_* \frac{\partial}{\partial s}, \nabla_{\partial/\partial t}\alpha_* \frac{\partial}{\partial t}\right) dt - \int_{t_{i-1}}^{t_i} \frac{d}{dt} g\left(\alpha_* \frac{\partial}{\partial s}, \alpha_* \frac{\partial}{\partial t}\right) dt.$$

Thus we have

$$L_i{}'(0) = \int_{t_{i-1}}^{t_i} g(V, c'')\, dt - [g(V, c')]_{t_{i-1}^+}^{t_i^-}$$

which implies that

$$L'(0) = \int_a^b g(V, c'')\, dt + \sum_{i=0}^{k} g(V(t_i), \Delta_{t_i}(c'))$$

as required. □

If $c : [a, b] \to M$ is a timelike geodesic, then $c'' = \nabla_{c'} c' = 0$ and $\Delta_{t_i}(c') = 0$ for all $i = 1, 2, \ldots, k-1$. Thus for a timelike geodesic, the first variation formula simplifies as follows.

Corollary 12.24. *If $c : [a, b] \to (M, g)$ is a unit speed timelike geodesic segment and $\alpha : [a, b] \times (-\epsilon, \epsilon) \to (M, g)$ is a variation of c, then*

$$L'(0) = -g(V, c')|_a^b.$$

Now let H be a spacelike hypersurface and assume that $c : [a, b] \to M$ is a unit speed timelike curve with $c(a) \in H$. In studying focal points of the submanifold H, attention may be restricted to variations $\alpha : [a, b] \times (-\epsilon, \epsilon) \to (M, g)$ of c which start on H [i.e., $\alpha(a, s) \in H$ for all $s \in (-\epsilon, \epsilon)$] and which end at $c(b)$ [i.e., $\alpha(b, s) = c(b)$ for all $s \in (-\epsilon, \epsilon)$]. For these variations, $V(b) = 0$ and $V(a)$ is tangent to H at $c(a)$. Proposition 12.23 then yields the following first variation formula:

$$(12.7) \qquad L'(0) = \int_a^b g(V, c'')\big|_{(t,0)}\, dt + \sum_{i=1}^{k-1} g(V(t_i), \Delta_{t_i}(c')) + g(V, c')\big|_{t=a}.$$

Given a spacelike hypersurface H without boundary in (M, g), fix a point $q \in M - H$. Consider the collection (possibly empty) of all timelike curves which join some point of H to q. If this collection contains a longest curve $c : [a, b] \to (M, g)$, then c must be a smooth timelike geodesic by the usual arguments that timelike geodesics locally maximize arc length and using the first variational formula to see that c has no corners. Thus assume that the unit speed timelike geodesic $c : [a, b] \to (M, g)$ is the longest curve from H to q. If $\alpha : [a, b] \times (-\epsilon, \epsilon) \to (M, g)$ is a variation of c through timelike curves with $\alpha(a, s) \in H$ and $\alpha(b, s) = q$ for all s with $-\epsilon \leq s \leq \epsilon$, then using $V(b) = 0$ and Corollary 12.24 we have

$$L'(0) = g(V(a), c'(a)).$$

On the other hand, $L'(0) = 0$ since c is of maximal length from H to q. Thus $V(a)$ is orthogonal to $c'(a)$. Since variations α as above may be constructed with $V(a) \in T_{c(a)}H$ arbitrary, it follows that c must be orthogonal to H at $c(a)$. Thus we have obtained the following standard result. Note also that it is necessary for H to have no boundary in order to be sure that the extremal c is perpendicular to H at $c(a)$.

Proposition 12.25. *Let H be a spacelike hypersurface without boundary in (M, g). Assume $c : [a, b] \to (M, g)$ is a timelike curve from H to the point $q = c(b) \notin H$ which is of maximal length among all timelike curves from H to q. Then c is a timelike geodesic segment, and c is orthogonal to H at $c(a)$.*

The variation vector field $V(t) = \alpha_*(\partial/\partial s)|_{(t,0)}$ of a variation α along c may have discontinuities in its derivative at the t-parameter values $\{t_1, t_2, \ldots, t_{k-1}\}$ at which α may fail to be smooth. Thus the normal component $N = \alpha_*\partial/\partial s + g(\alpha_*\partial/\partial s, \alpha_*\partial/\partial t)\alpha_*\partial/\partial t$ of V along c may also fail to be smooth at these parameter values. We now derive the second variation formula for $L''(0)$ in terms of N and V [cf. Bölts (1977, pp. 86–90), Hawking and Ellis (1973, p. 108)].

Proposition 12.26. *Let $c : [a, b] \to (M, g)$ be a unit speed timelike geodesic segment, and let $\alpha : [a, b] \times (-\epsilon, \epsilon) \to (M, g)$ be a piecewise smooth* *variation of c which is smooth on each set* $(t_{i-1}, t_i) \times (-\epsilon, \epsilon)$ *for the partition* $a = t_0 < t_1 < \cdots < t_k = b$ *of* $[a,b]$. *Let* $V(t) = \alpha_* \partial/\partial s|_{(t,0)}$ *denote the variation vector field of* α *along c, and set*

$$\begin{aligned} N(t) &= V(t) + g(V(t), c'(t))c'(t) \\ &= \alpha_* \left.\frac{\partial}{\partial s}\right|_{(t,0)} + g\left(\alpha_* \frac{\partial}{\partial s}, \alpha_* \frac{\partial}{\partial t}\right) \alpha_* \left.\frac{\partial}{\partial t}\right|_{(t,0)}. \end{aligned}$$

Then

$$L''(0) = \int_a^b g(N'' + R(V,c')c', N)\big|_{(t,0)}\, dt + \sum_{i=0}^{k} g(N(t_i), \Delta_{t_i}(N')) - g(\nabla_{\partial/\partial s} V, c')\big|_b^a.$$

Proof. Setting $L_i = L \,|\, [t_{i-1}, t_i]$ and recalling that

$$\frac{dL_i}{ds} = \int_{t_{i-1}}^{t_i} \left[-g\left(\alpha_* \frac{\partial}{\partial t}, \alpha_* \frac{\partial}{\partial t}\right)\right]^{-1/2} \left[-g\left(\nabla_{\partial/\partial t}\left(\alpha_* \frac{\partial}{\partial s}\right), \alpha_* \frac{\partial}{\partial t}\right)\right] dt,$$

we obtain

$$\frac{d^2 L_i}{ds^2} = \int_{t_{i-1}}^{t_i} \frac{d}{ds}\left\{\left[-g\left(\alpha_* \frac{\partial}{\partial t}, \alpha_* \frac{\partial}{\partial t}\right)\right]^{-1/2} \left[-g\left(\nabla_{\partial/\partial t}\left(\alpha_* \frac{\partial}{\partial s}\right), \alpha_* \frac{\partial}{\partial t}\right)\right]\right\} dt.$$

Differentiating the expression under the integral sign and using the identity $\nabla_{\partial/\partial s} \alpha_* \frac{\partial}{\partial t} - \nabla_{\partial/\partial t} \alpha_* \frac{\partial}{\partial s} = \alpha_* \left[\frac{\partial}{\partial s}, \frac{\partial}{\partial t}\right] = 0$ yields the integrand

(12.8)

$$\frac{-g(\nabla_{\partial/\partial s}\nabla_{\partial/\partial t}\alpha_*\partial/\partial s, \alpha_*\partial/\partial t) - g(\nabla_{\partial/\partial t}\alpha_*\partial/\partial s, \nabla_{\partial/\partial s}\alpha_*\partial/\partial t)}{[-g(\alpha_*\partial/\partial t, \alpha_*\partial/\partial t)]^{1/2}} + \frac{g(\nabla_{\partial/\partial t}\alpha_*\partial/\partial s, \alpha_*\partial/\partial t) g(\nabla_{\partial/\partial t}\alpha_*\partial/\partial s, \alpha_*\partial/\partial t)}{[-g(\alpha_*\partial/\partial t, \alpha_*\partial/\partial t)]^{1/2} g(\alpha_*\partial/\partial t, \alpha_*\partial/\partial t)}$$

for $(t,s) \in (t_{i-1}, t_i) \times (-\epsilon, \epsilon)$. Furthermore, we have

$$\begin{aligned} \nabla_{\partial/\partial t}\left[g\left(\alpha_* \frac{\partial}{\partial s}, \alpha_* \frac{\partial}{\partial t}\right) \alpha_* \frac{\partial}{\partial t}\right]\Bigg|_{(t,0)} &= \frac{d}{dt}\left[g\left(\alpha_* \frac{\partial}{\partial s}, \alpha_* \frac{\partial}{\partial t}\right)\right] \alpha_* \left.\frac{\partial}{\partial t}\right|_{(t,0)} \\ &\quad + g\left(\alpha_* \frac{\partial}{\partial s}, \alpha_* \frac{\partial}{\partial t}\right) \nabla_{\partial/\partial t} \alpha_* \left.\frac{\partial}{\partial t}\right|_{(t,0)} \\ &= \frac{d}{dt}\left[g\left(\alpha_* \frac{\partial}{\partial s}, \alpha_* \frac{\partial}{\partial t}\right)\right] \alpha_* \left.\frac{\partial}{\partial t}\right|_{(t,0)} \end{aligned}$$

since $\nabla_{\partial/\partial t}\alpha_*\partial/\partial t\big|_{(t,0)} = \nabla_{\partial/\partial t}c'(t) = 0$, as c is a smooth geodesic. Also

$$g\left(\nabla_{\partial/\partial t}N, \alpha_*\frac{\partial}{\partial t}\right)\bigg|_{(t,0)} = \frac{d}{dt}\left(g\left(N, \alpha_*\frac{\partial}{\partial t}\right)\right)\bigg|_{(t,0)} - g\left(N, \nabla_{\partial/\partial t}\alpha_*\frac{\partial}{\partial t}\right)\bigg|_{(t,0)} = 0$$

since c is a geodesic and $g(N, \alpha_*\partial/\partial t)|_{(t,0)} = 0$ for all $t \in (t_{i-1}, t_i)$. Using these last two calculations, we then obtain

$$\begin{aligned} &g\left(\nabla_{\partial/\partial t}N, \nabla_{\partial/\partial t}\left[g\left(\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\alpha_*\frac{\partial}{\partial t}\right]\right)\bigg|_{(t,0)} \\ &= \frac{d}{dt}\left(g\left(\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\right)\bigg|_{(t,0)} g\left(\nabla_{\partial/\partial t}N, \alpha_*\frac{\partial}{\partial t}\right)\bigg|_{(t,0)} = 0. \end{aligned}$$

Thus since

$$\alpha_*\frac{\partial}{\partial s}\bigg|_{(t,0)} = N(t) - g\left(\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\alpha_*\frac{\partial}{\partial t}\bigg|_{(t,0)},$$

it follows from these calculations that

$$\begin{aligned} &g\left(\nabla_{\partial/\partial t}\alpha_*\frac{\partial}{\partial s}, \nabla_{\partial/\partial s}\alpha_*\frac{\partial}{\partial t}\right)\bigg|_{(t,0)} \\ &= g\left(\nabla_{\partial/\partial t}\alpha_*\frac{\partial}{\partial s}, \nabla_{\partial/\partial t}\alpha_*\frac{\partial}{\partial s}\right)\bigg|_{(t,0)} \\ &= g\left(\nabla_{\partial/\partial t}N, \nabla_{\partial/\partial t}N\right)\big|_{(t,0)} \\ &\quad - 2g\left(\nabla_{\partial/\partial t}N, \nabla_{\partial/\partial t}\left[g\left(\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\alpha_*\frac{\partial}{\partial t}\right]\right)\bigg|_{(t,0)} \\ &\quad + g\left(\nabla_{\partial/\partial t}\left[g\left(\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\alpha_*\frac{\partial}{\partial t}\right], \nabla_{\partial/\partial t}\left[g\left(\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\alpha_*\frac{\partial}{\partial t}\right]\right)\bigg|_{(t,0)} \\ &= g\left(\nabla_{\partial/\partial t}N, \nabla_{\partial/\partial t}N\right)\big|_{(t,0)} \\ &\quad + \left\{\left[\frac{d}{dt}g\left(\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\right]^2 g\left(\alpha_*\frac{\partial}{\partial t}, \alpha_*\frac{\partial}{\partial t}\right)\right\}\bigg|_{(t,0)} \\ &= g\left(\nabla_{\partial/\partial t}N, \nabla_{\partial/\partial t}N\right)\big|_{(t,0)} - \left[g\left(\nabla_{\partial/\partial t}\left(\alpha_*\frac{\partial}{\partial s}\right), \alpha_*\frac{\partial}{\partial t}\right)\right]^2\bigg|_{(t,0)}. \end{aligned}$$

Substituting this expression for $g(\nabla_{\partial/\partial t}\alpha_*\partial/\partial s, \nabla_{\partial/\partial s}\alpha_*\partial/\partial t)|_{(t,0)}$ in (12.8) and using $g(\alpha_*\partial/\partial t, \alpha_*\partial/\partial t)|_{(t,0)} = -1$, we obtain

$$\frac{d}{ds}\left\{\left[-g\left(\alpha_*\frac{\partial}{\partial t}, \alpha_*\frac{\partial}{\partial t}\right)\right]^{-1/2}\left[-g\left(\nabla_{\partial/\partial t}\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\right]\right\}\Bigg|_{(t,0)}$$

$$= -g\left(\nabla_{\partial/\partial s}\nabla_{\partial/\partial t}\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\bigg|_{(t,0)} - g\left(\nabla_{\partial/\partial t}N, \nabla_{\partial/\partial t}N\right)\big|_{(t,0)}.$$

This yields

$$L_i''(0) =$$

$$\int_{t_{i-1}}^{t_i}\left[-g\left(\nabla_{\partial/\partial s}\nabla_{\partial/\partial t}\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\bigg|_{(t,0)} - g\left(\nabla_{\partial/\partial t}N, \nabla_{\partial/\partial t}N\right)\big|_{(t,0)}\right]dt.$$

Also

$$R\left(\alpha_*\frac{\partial}{\partial t}, \alpha_*\frac{\partial}{\partial s}\right)\alpha_*\frac{\partial}{\partial s} = \nabla_{\partial/\partial t}\nabla_{\partial/\partial s}\left(\alpha_*\frac{\partial}{\partial s}\right) - \nabla_{\partial/\partial s}\nabla_{\partial/\partial t}\left(\alpha_*\frac{\partial}{\partial s}\right)$$

since

$$\left[\alpha_*\frac{\partial}{\partial t}, \alpha_*\frac{\partial}{\partial s}\right] = \alpha_*\left[\frac{\partial}{\partial t}, \frac{\partial}{\partial s}\right] = 0.$$

Since $c(t) = \alpha(t, 0)$ is a geodesic,

$$g\left(\nabla_{\partial/\partial t}\nabla_{\partial/\partial s}\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\bigg|_{(t,0)} = \frac{d}{dt}g\left(\nabla_{\partial/\partial s}\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\bigg|_{(t,0)}.$$

Therefore,

$$L_i''(0) = \int_{t_{i-1}}^{t_i}\left[g\left(R\left(\alpha_*\frac{\partial}{\partial t}, \alpha_*\frac{\partial}{\partial s}\right)\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\bigg|_{(t,0)}\right.$$

$$\left. - g\left(\nabla_{\partial/\partial t}N, \nabla_{\partial/\partial t}N\right)\big|_{(t,0)}\right]dt$$

$$- g\left(\nabla_{\partial/\partial s}\alpha_*\frac{\partial}{\partial s}, \alpha_*\frac{\partial}{\partial t}\right)\bigg|_{t_{i-1}}^{t_i}.$$

The equations

$$-g(\nabla_{\partial/\partial t}N, \nabla_{\partial/\partial t}N)\big|_{(t,0)}$$

$$= -\frac{d}{dt}g(N, \nabla_{\partial/\partial t}N)\bigg|_{(t,0)} + g(N, \nabla_{\partial/\partial t}\nabla_{\partial/\partial t}N)|_{(t,0)}$$

and $V(t) = \alpha_* \partial/\partial s|_{(t,0)}$ together with $L = \sum_{i=1}^{k} L_i$ yield

$$\begin{aligned} L''(0) = &\int_a^b \left[g(R(c', V)V, c')\big|_{(t,0)} + g(N, N'')\big|_{(t,0)} \right] dt \\ &- \sum_{i=1}^{k} g(N, N')\big|_{t_{i-1}^+}^{t_i^-} - g(\nabla_{\partial/\partial s} V, c')\big|_a^b \\ = &\int_a^b \left[g(R(V, c')c', N - g(V, c')c')\big|_t + g(N, N'')\big|_t \right] dt \\ &+ \sum_{i=0}^{k} g(N, \Delta_{t_i}(N')) - g(\nabla_{\partial/\partial s} V, c')\big|_a^b \end{aligned}$$

since $V = N - g(V, c')c'$. Thus

$$\begin{aligned} L''(0) = &\int_a^b g(N'' + R(V, c')c', N)\big|_t \, dt \\ &+ \sum_{i=0}^{k} g(N(t_i), \Delta_{t_i}(N')) - g(\nabla_{\partial/\partial s} V, c')\big|_a^b \end{aligned}$$

where

$$\nabla_{\partial/\partial s} V\big|_{(t,0)} = \nabla_{\partial/\partial s} \alpha_* \partial/\partial s\big|_{(t,0)} \text{ as required.} \quad \square$$

Proposition 12.26 has the following consequence.

Corollary 12.27. *Let H be a spacelike hypersurface, and assume that $c : [a, b] \to (M, g)$ is a unit speed timelike geodesic which is orthogonal to H at $p = c(a) \in H$. Suppose that $\alpha : [a, b] \times (-\epsilon, \epsilon) \to (M, g)$ is a variation of c such that $\alpha(a, s) \in H$ and $\alpha(b, s) = q = c(b)$ for all s with $-\epsilon \leq s \leq \epsilon$. If $V = \alpha_* \partial/\partial s|_{(t,0)}$ and $N = V + g(V, c')c'$, then*

$$\begin{aligned} L''(0) = &\int_a^b g(N'' + R(V, c')c', N)\big|_t \, dt + \sum_{i=1}^{k-1} g(N(t_i), \Delta_{t_i}(N')) \\ &+ g(N, N')\big|_p + g(L_{c'}(N), N)\big|_p \end{aligned}$$

where $L_{c'}$ is the second fundamental form operator of H at p.

Proof. In view of Proposition 12.26 and the equation

$$\sum_{i=0}^{k} g(N(t_i), \Delta_{t_i}(N')) = \sum_{i=1}^{k-1} g(N(t_i), \Delta_{t_i}(N')) + g(N, N')\big|_p,$$

it is only necessary to show that $-g(\nabla_{\partial/\partial s}V, c')|_a^b$ is equal to $g(L_{c'}(N), N)$. To this end, we first note that $\alpha(b,s) = q$ implies that $\alpha_*\partial/\partial s|_{(b,s)} = 0$ for all s which yields $g(\nabla_{\partial/\partial s}V, c')|_b = 0$. Also $\alpha(a,s) \in H$ for all s with $-\epsilon < s < \epsilon$ implies that $\alpha_*\partial/\partial s|_{(a,s)}$ is tangential to H for all s, and hence $N(a) = \alpha_*\partial/\partial s|_{(a,0)}$. Let $\gamma(s) = \alpha(a,s)$ for all s with $-\epsilon < s < \epsilon$. Extend the vector $N(a) \in T_pH$ to a local vector field X along H with $X \circ \gamma(s) = \alpha_*\partial/\partial s|_{(a,s)}$ for all s with $-\epsilon < s < \epsilon$. Then

$$g(L_{c'(a)}(N), N) = g(\nabla_{X(a)}X, c'(a))$$

by Definition 3.48. Also let η be a unit normal field to H near p with $\eta(p) = c'(a)$. Then we have

$$\begin{aligned} g(\nabla_{\partial/\partial s}V, c'(a)) &= g\left(\nabla_{\partial/\partial s}\alpha_*\frac{\partial}{\partial s}, \eta\circ\alpha\right)\bigg|_{(a,0)} \\ &= \frac{d}{ds}\left(g\left(\alpha_*\frac{\partial}{\partial s}, \eta\circ\alpha\right)\right)\bigg|_{(a,0)} \\ &\qquad - g\left(\alpha_*\frac{\partial}{\partial s}, \nabla_{\partial/\partial s}\eta\circ\alpha\right)\bigg|_{(a,0)}. \end{aligned}$$

But $g(\alpha_*\partial/\partial s, \eta\circ\alpha)|_{(a,s)} = 0$ for all s since $\alpha_*\partial/\partial s|_{(a,s)}$ is tangential to H. Thus we obtain

$$\begin{aligned} g(\nabla_{\partial/\partial s}V, c'(a)) &= -g\left(\alpha_*\frac{\partial}{\partial s}, \nabla_{\partial/\partial s}\eta\circ\alpha\right)\bigg|_{(a,0)} \\ &= -g(N(a), \nabla_{N(a)}\eta) \\ &= -g(X, \nabla_X\eta)|_{c(a)} \\ &= -X|_p(g(X,\eta)) + g(\nabla_X X|_p, \eta(a)) \\ &= 0 + g(\nabla_X X|_p, c'(a)) \\ &= g(L_{c'(a)}(N), N) \end{aligned}$$

as required. Here we have used the fact that since $X|_p = \alpha_*\partial/\partial s|_{(a,0)}$, $X|_p(g(X,\eta))$ may be calculated as

$$X|_p(g(X,\eta)) = \frac{d}{ds}(g(X\circ\gamma(s), \eta(s)))\bigg|_{s=0} = \frac{d}{ds}(0) = 0. \quad \square$$

In view of Corollary 12.27, the index $I_H(V, V)$ of a vector field V along a timelike geodesic orthogonal to a spacelike hypersurface should be defined as follows [cf. Bölts (1977, p. 94)].

Definition 12.28. (*Spacelike Hypersurface Index Form*) Let $c : [a, b] \to (M, g)$ be a unit speed timelike geodesic which is orthogonal to a spacelike hypersurface H at $c(a)$. Assume that Z is a piecewise smooth vector field along c which is orthogonal to c. If $Z(a) \neq 0$ and $Z(b) = 0$, then the *index of Z with respect to H* is given by

$$I_H(Z, Z) = I(Z, Z) + g(L_{c'}(Z), Z)|_a$$

where

$$I(Z, Z) = \int_a^b g(Z'' + R(Z, c')c', Z)|_t \, dt + \sum_{i=0}^{k-1} g(Z(t_i), \Delta_{t_i}(Z')),$$

where a partition $\{t_i\}$ of $[a, b]$ is chosen such that Z is differentiable except at the t_i's.

Let $\alpha : [a, b] \times (-\epsilon, \epsilon) \to (M, g)$ be a variation of the timelike geodesic c, and assume the variation vector field $V = \alpha_* \partial/\partial s|_{(t,0)}$ satisfies the conditions of Definition 12.28. Then equation (10.4) of Section 10.1 together with Corollary 12.27 yield $L''(0) = I_H(V, V)$. This may also be written

$$L_{**}(V) = I_H(V, V).$$

Using the index form I_H, it may now be shown that a timelike geodesic orthogonal to a spacelike hypersurface H fails to maximize arc length to H after the first focal point. Recall that $V^\perp(c)$ consists of piecewise smooth vector fields along c which are orthogonal to c.

Proposition 12.29. *Let $c : [a, b] \to M$ be a unit speed timelike geodesic orthogonal to a spacelike hypersurface H (without boundary) at a point $p = c(a) \in H$. If for some $t_k \in (a, b)$ the point $c(t_k)$ is a focal point to H along c, then there is a variation vector field $Z \in V^\perp(c)$ with $Z(a)$ tangential to H*

and $Z(b) = 0$ such that $I_H(Z,Z) > 0$. Consequently, there are timelike curves from H to $c(b)$ which are longer than c.

Proof. By hypothesis there exists a nontrivial Jacobi field J_1 along c with J_1 orthogonal to c, $J_1(t_k) = 0$ and $J_1{}'(a) = -L_{c'(a)}J_1(a)$. Define a piecewise smooth Jacobi field J along c by

$$J(t) = \begin{cases} J_1(t) & \text{if } a \le t \le t_k, \\ 0 & \text{if } t_k \le t \le b. \end{cases}$$

Since $\Delta_{t_k}(J') \ne 0$, we may construct a smooth vector field V orthogonal to c such that $V'(a) = V(a) = V(b) = 0$ and $g(V(t_k), \Delta_{t_k}(J')) = -1$. Define a vector field Z in $V^\perp(c)$ by

$$Z = \frac{1}{r}J - rV \qquad \text{for } r \in \mathbb{R} - \{0\}.$$

Then $Z'(a) = -L_{c'(a)}Z(a)$, and the index $I_H(Z,Z)$ is given by

$$\begin{aligned} I_H(Z,Z) &= I(Z,Z) + g(L_{c'}(Z), Z)\big|_a \\ &= I(Z,Z) + g(L_{c'}(r^{-1}J - rV), r^{-1}J - rV)\big|_a \\ &= I(Z,Z) + r^{-2}g(L_{c'}J, J)\big|_a \\ &= I(Z,Z) + r^{-2}g(-J', J)\big|_a \\ &= r^{-2}I(J,J) + r^2 I(V,V) - 2I(J,V) + r^{-2}g(-J', J)\big|_a \\ &= r^2 I(V,V) - 2I(J,V). \end{aligned}$$

Here we have used

$$I(J,J) = \sum_{i=0}^{k-1} g(\Delta_{t_i}(J'), J(t_i)) = g(J', J)\big|_a.$$

Since J is a piecewise smooth Jacobi field, equation (10.4) of Section 10.1 implies that

$$I(J,V) = g(V(t_k), \Delta_{t_k}(J')) = -1.$$

Consequently,

$$I_H(Z,Z) = r^2 I(V,V) + 2$$

which shows that for sufficiently small $r \neq 0$ the index satisfies

$$I_H(Z,Z) > 0.$$

This last inequality and the condition $Z'(a) = -L_{c'(a)}(Z(a))$ imply that there exist small variations of c with variation vector field Z which join H to $c(b)$ and have length greater than c. □

We now turn our attention to focal points along null geodesics which are orthogonal to $(n-2)$-dimensional spacelike submanifolds. If H is a spacelike $(n-2)$-dimensional submanifold, then the induced metric on T_pH is positive definite and the induced metric on $T_p^{\perp}H$ is a two-dimensional Minkowski metric for each $p \in H$. Thus there are exactly two null lines through the origin in $T_p^{\perp}H$. Since the time orientation of (M,g) induces a time-orientation on $T_p^{\perp}H$, there are thus two well-defined future null directions in $T_p^{\perp}H$ for each $p \in H$. Locally, we may then choose a smooth pseudo-orthonormal basis of vector fields $E_1, E_2, \ldots, E_{n-1}, E_n$ on H such that E_{n-1} and E_n are future directed null vectors in $T_p^{\perp}H$ for $p \in H$. That is,

$$\begin{aligned}
&g(E_i, E_j) = \delta_{ij} && \text{if } 1 \leq i,j \leq n-2,\\
&g(E_i, E_{n-1}) = g(E_i, E_n) = 0 && \text{if } 1 \leq i \leq n-2,\\
&g(E_n, E_n) = g(E_{n-1}, E_{n-1}) = 0,\\
\text{and}\quad &g(E_n, E_{n-1}) = -1.
\end{aligned}$$

The null vector fields E_{n-1} and E_n defined locally on H give rise to second fundamental forms $L_{E_{n-1}}$ and L_{E_n}, respectively, which are locally defined $(1,1)$ tensor fields on H.

If $\beta : [a,b] \to (M,g)$ is a future directed null geodesic with $\beta'(a) = E_n(p)$ [or $\beta'(a) = E_{n-1}(p)$] at $p \in H$, the tangent vectors $E_1(p), E_2(p), \ldots, E_n(p)$ at p may then be parallel translated along β to give a pseudo-orthonormal basis along β which will also be denoted by $E_1, E_2, \ldots, E_n$. The set of vectors normal to $\beta'(t)$ is thus the space $N(\beta(t))$ spanned by $E_1, E_2, \ldots, E_{n-2}, \beta'$, and we may form the quotient space $G(\beta(t)) = N(\beta(t))/[\beta'(t)]$ with corresponding quotient bundle $G(\beta)$ as in Section 10.3. If $\pi : N(\beta) \to G(\beta)$ denotes the projection map, then $\pi\,|_{T_pH} : T_pH \to G(\beta(a))$ is a vector space isomorphism. Hence we may project the second fundamental form operator $L_{E_i} : T_pH \to T_pH$ to an operator $\overline{L}_{E_i} : G(\beta(a)) \to G(\beta(a))$ by setting

$$\overline{L}_{E_i} = \pi \circ L_{E_i} \circ (\pi \,|\, T_pH)^{-1} \qquad \text{for } i = n-1, n.$$

Let $\alpha : [a,b] \times (-\epsilon, \epsilon) \to (M,g)$ be a variation of the null geodesic β such that $\alpha(a,s) \in H$ for all s with $-\epsilon < s < \epsilon$. If $V = \alpha_* \partial/\partial s$ and $W = \alpha_* \partial/\partial t$, we will also require that $W(a,s) = E_n(\alpha(a,s))$ for all s with $-\epsilon < s < \epsilon$. Thus the neighboring curve $\alpha(\,\cdot\,, s)$ starts at $\alpha(a,s)$ on H perpendicular to H and has as initial direction the null vector $E_n(\alpha(a,s))$ for all s.

We now calculate $V'(a)$.

Lemma 12.30. *Let $V = \alpha_* \partial/\partial s$ be tangential to H for $t = a$ and s arbitrary as above. Then $V'(a) = -L_{\beta'(a)}(V(a)) + \lambda \beta'(a)$ for some constant $\lambda \in \mathbb{R}$.*

Proof. Using $[V, W] = [\alpha_* \partial/\partial s, \alpha_* \partial/\partial t] = \alpha_*[\partial/\partial s, \partial/\partial t] = 0$, we have $\nabla_V W = \nabla_W V = \nabla_{\beta'} V$ at $t = a$, $s = 0$.

On the other hand, $W(a,s) = E_n(\alpha(a,s))$ and $g(E_n, E_n) = 0$ yield $0 = V(g(W,W)) = 2g(\nabla_V W, W) = 2g(\nabla_W V, W) = 2g(\nabla_{\beta'} V, \beta')$ at $t = a$, $s = 0$. Thus

$$\nabla_{\beta'(a)} V \in N(\beta(a)) = T_pH \oplus [\beta'(a)].$$

Since $L_{\beta'(a)}(V(a)) \in T_{\beta(a)}H$ and $\nabla_{\beta'(a)} V \in N(\beta(a))$, it thus suffices to show that

$$g(L_{\beta'(a)}(V(a)), y) = -g(\nabla_{\beta'(a)} V, y)$$

for every $y \in T_{\beta(a)}H$ to establish the result. To calculate $g(L_{\beta'(a)}(V(a)), y)$, extend y to a local vector field Y along the curve $s \to \alpha(a,s)$ that is tangent to H. Then we have

$$\begin{aligned} g(L_{\beta'(a)}(V(a)), y) &= g\left(\nabla_V Y\big|_{(a,0)}, \beta'(a)\right) \\ &= g(\nabla_V Y, W)\big|_{(a,0)} \\ &= \alpha_* \left.\frac{\partial}{\partial s}\right|_{(a,0)} (g(Y,W)) - g(Y, \nabla_V W)\big|_{(a,0)} \\ &= 0 - g(Y, \nabla_W V)\big|_{(a,0)} \\ &= -g(y, \nabla_{\beta'(a)} V) \end{aligned}$$

where we have used $\alpha_* \partial/\partial s|_{(a,0)}(g(Y,W)) = 0$ since $g(Y,W) = g(Y, E_n) = 0$ along the curve $s \to \alpha(a,s)$. □

If V is to be a Jacobi field measuring the separation of a congruence of null geodesics perpendicular to H, then the last result implies that the vector class $\overline{V} = \pi(V)$ along β should satisfy the initial condition

$$\overline{V}'(a) = -\overline{L}_{\beta'(a)}\overline{V}(a)$$

at $p = \beta(a) \in H$. Here $\overline{L}_{\beta'} = \pi \circ L_{\beta'} \circ (\pi \,|\, T_p H)^{-1}$ as above. This motivates the following definition of focal point along a null geodesic perpendicular to H [cf. Hawking and Ellis (1973, p. 102), Bölts (1977, p. 56)].

Definition 12.31. (*Focal Point on a Null Geodesic*) Let H be a spacelike submanifold of dimension $(n-2)$, and suppose that $\beta : [a,b] \to (M,g)$ is a null geodesic orthogonal to H at $p = \beta(a)$. Then $t_0 \in (a,b]$ is said to be a *focal point of H along β* if there is a nontrivial smooth Jacobi class $\overline{J}$ in $G(\beta)$ such that $\overline{J}'(a) = -\overline{L}_{\beta'(a)}\overline{J}(a)$ and $\overline{J}(t_0) = [\beta'(t_0)]$.

We noted above that a timelike geodesic orthogonal to a spacelike hypersurface fails to be the longest nonspacelike curve to the hypersurface after the first focal point (cf. Proposition 12.29). A similar result holds for a null geodesic which is orthogonal to an $(n-2)$-dimensional spacelike submanifold [cf. Hawking and Ellis (1973, p. 116), Bölts (1977, p. 123)]. We state this result as the following proposition.

Proposition 12.32. *Let H be a spacelike submanifold of (M, g) of dimension $n-2$, and let $\beta : [a, b] \to (M, g)$ be a null geodesic orthogonal at $\beta(a)$ to H. If $t_0 \in (a, b)$ is a focal point of H along β, then there is a timelike curve from H to $\beta(b)$. Thus β does not maximize the distance to H after the first focal point.*

A simple example of a focal point is shown in Figure 12.3.

Using the same type of reasoning as in Proposition 12.22, the following result may also be established.

Proposition 12.33. *Let (M, g) be a space–time of dimension $n \geq 3$, and let H be a spacelike submanifold of dimension $n-2$. Suppose that $\beta : J \to (M, g)$ is an inextendible null geodesic which is orthogonal to H at $p = \beta(t_1)$ and satisfies the curvature condition $\mathrm{Ric}(\beta'(t), \beta'(t)) \geq 0$ for all $t \in J$. If $-\operatorname{tr}(L_{\beta'(t_1)})$ has the negative [respectively, positive] value θ_1 at p, then there is a focal point t_0 to H along β in the parameter interval from t_1 to $t_1-(n-2)/\theta_1$ [respectively, in the parameter interval from $t_1-(n-2)/\theta_1$ to t_1] provided that $t_0 \in J$.*

A particularly important case occurs when H is a compact spacelike $(n-2)$-dimensional submanifold which satisfies the condition $(\operatorname{tr} L_{E_n}) \cdot (\operatorname{tr} L_{E_{n-1}}) > 0$ at each point [cf. Hawking and Ellis (1973, p. 262)]. Here if $\{e_1, e_2, \ldots, e_{n-2}\}$ is an orthonormal basis for T_pH, then $\operatorname{tr} L_{E_n}$ may be calculated as

$$\operatorname{tr} L_{E_n} = \sum_{i=1}^{n-2} g(L_{E_n}(e_i), e_i).$$

Definition 12.34. (*Closed Trapped Surface*) Suppose that H is a compact spacelike submanifold without boundary of (M, g) of dimension $n-2$. Let E_n and E_{n-1} be linearly independent future directed null vector fields on H as above. Assume that L_1 and L_2 are the second fundamental forms on H corresponding to E_n and E_{n-1}, respectively. Then H is said to be a *closed trapped surface* if $\operatorname{tr} L_1$ and $\operatorname{tr} L_2$ are both either always positive or always negative on H.

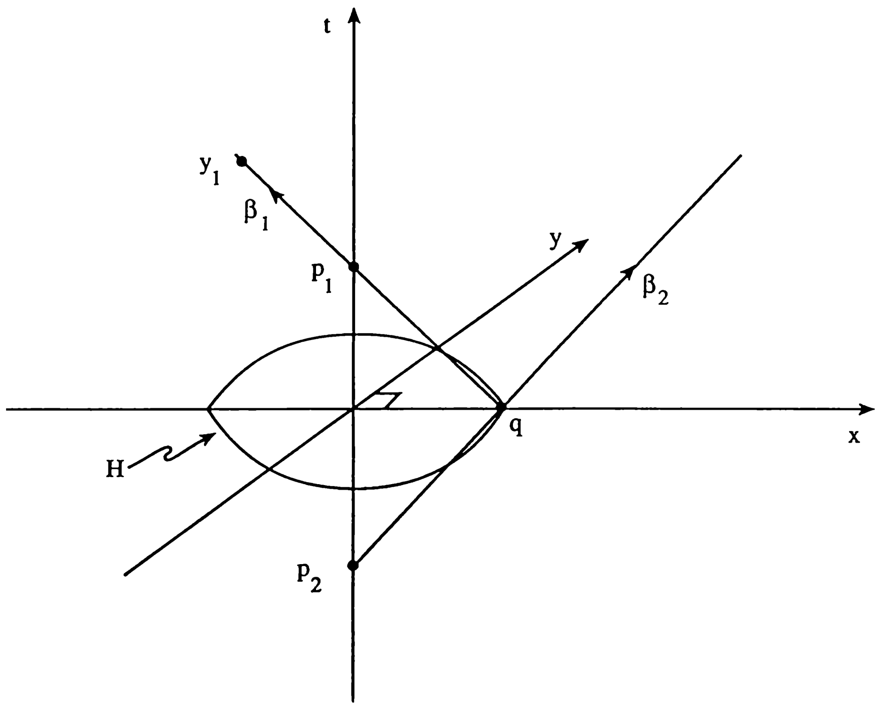


FIGURE 12.3. Let (M, g) be three-dimensional Minkowski space–time with the usual metric $ds^2 = -dt^2 + dx^2 + dy^2$. Let H be a circle of radius a in the x–y plane. Then H is a spacelike submanifold of codimension two. For each $q \in H$, there are exactly two future directed null geodesics β_1 and β_2 through q which are orthogonal to H. The focal point of H along β_1 is $p_1 = (0, 0, a)$, and the focal point along β_2 is $p_2 = (0, 0, -a)$. The point y_1 lies on β_1 beyond p_1. All of $H - \{q\}$ is contained in the chronological past of y_1. Thus there are timelike curves from points on H arbitrarily close to q to y_1, and $\beta_1[q, y_1]$ does not realize the distance from H to y_1.

A related concept is that of a trapped set [cf. Hawking and Penrose (1970, pp. 534–537)]. Recall that the future horismos of a set A is defined by $E^+(A) = J^+(A) - I^+(A)$.

Definition 12.35. (*Trapped Set*) A nonempty achronal set A is said to be *future* [respectively, *past*] *trapped* if $E^+(A)$ [respectively, $E^-(A)$] is compact. A *trapped set* is a set which is either future trapped or past trapped.

In general, a closed trapped surface need not be a trapped set and vice versa. However, in Proposition 12.45 we will show that under certain conditions, the existence of a closed trapped surface implies either null incompleteness or the existence of a trapped set. In establishing Proposition 12.45, we will need the following corollary to Proposition 12.33.

Corollary 12.36. *Let (M, g) be a space–time of dimension $n \geq 3$ which satisfies the condition* $\mathrm{Ric}(v, v) \geq 0$ *for all null vectors $v \in TM$. If (M, g) contains a closed trapped surface H, then either (1) or (2) or both is true:*

(1) *At least one of the sets $E^+(H)$ or $E^-(H)$ is compact.*

(2) *(M, g) is null incomplete.*

Proof. Assume that (M, g) is null complete and that $\operatorname{tr} L_1 > 0$ and $\operatorname{tr} L_2 > 0$ for all $q \in H$. Consider all future directed null geodesics which start at some point of H and have initial direction either E_{n-1} or E_n at this point. Each such geodesic contains a geodesic segment which goes from a point $q \in H$ to a first focal point p to H. Using Proposition 12.33 and the compactness of H, it follows that the union of all such null geodesic segments from H to a focal point is contained in a compact set K consisting of null geodesic segments starting on H. Now if $r \in E^+(H)$, then r can be joined to H by a past directed null geodesic but not by a past directed timelike curve. Thus $r \in K$. Hence $E^+(H) \subseteq K$. To show that $E^+(H)$ is closed, let $\{x_n\}$ be a sequence of points of $E^+(H)$. This sequence has a limit point $x \in K$. From the definition of K we have $x \in J^+(H)$. If $x \in I^+(H)$, then the open set $I^+(H)$ must contain some elements of the sequence $\{x_n\}$, contradicting $x_n \in E^+(H)$ for all n. Thus $x \notin I^+(H)$ which yields $x \in E^+(H)$. This shows that $E^+(H)$ is a closed subset of the compact set K and hence is compact.

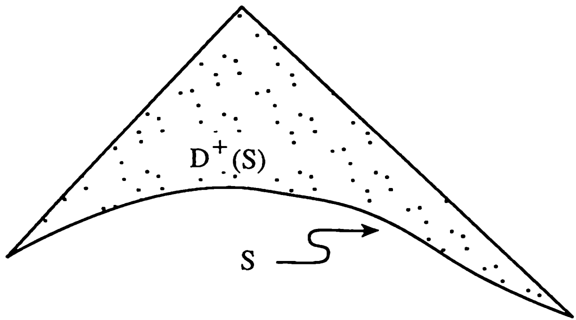


FIGURE 12.4. The future Cauchy development $D^+(S)$ of S consists of all points p such that every past inextendible nonspacelike curve through p intersects S.

If we assume that (M, g) is null complete and that $\operatorname{tr} L_1 < 0$ and $\operatorname{tr} L_2 < 0$ for all $q \in H$, then similar arguments show $E^-(H)$ is compact. This establishes the corollary. □

We now define the Cauchy developments of a closed subset S of (M, g) [cf. Hawking and Ellis (1973, pp. 201–205)].

Definition 12.37. (*Cauchy Development*) Let S be a closed subset of (M, g). The *future* [respectively, *past*] *Cauchy development* $D^+(S)$ [respectively, $D^-(S)$] consists of all points $p \in M$ such that every past [respectively, future] inextendible nonspacelike curve through p intersects S. The *Cauchy development* of S is $D(S) = D^+(S) \cup D^-(S)$ (cf. Figure 12.4).

Closed achronal sets have played an important role in causality theory and singularity theory in general relativity. They have the following property [cf. Hawking and Ellis (1973, pp. 209, 268), Hawking and Penrose (1970, p. 537)].

Proposition 12.38. *If S is a closed achronal set in the space–time (M, g), then $\operatorname{Int}(D(S)) = D(S) - \partial D(S)$ is globally hyperbolic.*

12.4 The Existence of Singularities

In this section we give the proofs of several singularity theorems in general relativity. In particular, we prove the main theorem of Hawking and Penrose (1970, p. 538). Our approach is somewhat different from Hawking and Penrose (1970) in that we show (M, g) is causally disconnected if it contains a trapped set and then apply Theorem 6.3 of Beem and Ehrlich (1979a, p. 172).

The basic technique in proving nonspacelike incompleteness is to first use physical or geometric assumptions on (M, g) to construct an inextendible non-spacelike geodesic which is maximal and hence contains no conjugate points. If (M, g) has dimension $n \geq 3$ and satisfies the generic condition and time-like convergence condition, this geodesic must then be incomplete by Theorem 12.18.

We first show that a chronological space–time with a sufficient number of conjugate points is strongly causal [cf. Hawking and Penrose (1970, p. 536), Lerner (1972, p. 41)].

Proposition 12.39. *If (M, g) is a chronological space–time such that each inextendible null geodesic has a pair of conjugate points, then (M, g) is strongly causal.*

Proof. Assume that strong causality fails at $p \in M$. Let U be a convex normal neighborhood of p, and let $V_k \subseteq U$ be a sequence of neighborhoods which converge to p. Since strong causality fails at p, for each k there is a future directed nonspacelike curve γ_k which starts in V_k, leaves U, and returns to V_k. Using Proposition 3.31, one may obtain an inextendible nonspacelike limit curve γ through p of the sequence $\{\gamma_k\}$. No two points of γ are chronologically related since otherwise one could obtain a closed timelike curve, contradicting the fact that (M, g) is chronological. Thus γ is a null geodesic. This yields a contradiction since by hypothesis each null geodesic has conjugate points and thus contains points which may be joined by timelike curves. $\square$

Proposition 12.39 and Theorem 12.18 then imply the following result.

Proposition 12.40. *Let (M, g) be a chronological space–time of dimension $n \geq 3$ which satisfies the generic condition and the timelike convergence condition. Then (M, g) is either strongly causal or null incomplete.*

We may now prove the following singularity theorem. The concept of a causally disconnected space–time has been given in Section 8.3, Definition 8.11.

Theorem 12.41. *Let (M, g) be a chronological space–time of dimension $n \geq 3$ which is causally disconnected. If (M, g) satisfies the generic condition and the timelike convergence condition, then (M, g) is nonspacelike incomplete.*

Proof. Assume all nonspacelike geodesics of (M, g) are complete. By Proposition 12.40, the space–time (M, g) is strongly causal, and by Theorem 12.18 every nonspacelike geodesic has conjugate points. On the other hand, Theorem 8.13 yields the existence of an inextendible maximal nonspacelike geodesic. But then this geodesic has no conjugate points, in contradiction. □

Recall that the *future* [respectively, *past*] *horismos* of a subset S of (M, g) is given by $E^+(S) = J^+(S) - I^+(S)$ [respectively, $E^-(S) = J^-(S) - I^-(S)$]. The achronal set S is said to be *future trapped* [respectively, *past trapped*] if $E^+(S)$ [respectively, $E^-(S)$] is compact. We now give conditions under which the existence of a trapped set implies causal disconnection.

Proposition 12.42. *Let (M, g) be a chronological space–time of dimension $n \geq 3$ such that each inextendible null geodesic has a pair of conjugate points. If (M, g) contains a future [respectively, past] trapped set S, then (M, g) is causally disconnected by $E^+(S)$ [respectively, $E^-(S)$].*

Proof. First, Proposition 12.39 shows that (M, g) is strongly causal. If S is assumed to be future trapped, then Corollary 8.16 yields a future inextendible timelike curve γ in the future Cauchy development $D^+(E^+(S))$. Extend γ to a future and past inextendible timelike curve in (M, g), still denoted by γ. Then γ intersects the achronal set $E^+(S)$ in exactly one point r. As in the proof of Proposition 8.18, we choose two sequences $\{p_n\}$ and $\{q_n\}$ on γ which diverge to infinity and satisfy $p_n \ll r \ll q_n$ for each n. To show that $\{p_n\}$,

$\{q_n\}$, and $E^+(S)$ causally disconnect (M,g), we must show that for each n, every nonspacelike curve $\lambda : [0,1] \to M$ with $\lambda(0) = p_n$ and $\lambda(1) = q_n$ meets $E^+(S)$. Given λ, extend λ to a past inextendible curve $\widetilde{\lambda}$ by traversing γ up to p_n and then traversing λ from p_n to q_n. As $q_n \in D^+(E^+(S))$, the curve $\widetilde{\lambda}$ must intersect $E^+(S)$. Since γ meets $E^+(S)$ only at r, it follows that λ intersects $E^+(S)$. This establishes the proposition if S is future trapped. A similar argument may be used if S is past trapped. □

Theorem 8.13, Proposition 12.39, and Proposition 12.42 now imply the main theorem of Hawking and Penrose (1970, p. 538).

Theorem 12.43. *No space–time (M,g) of dimension $n \geq 3$ can satisfy all of the following three requirements together:*

(1) *(M,g) contains no closed timelike curves.*
(2) *Every inextendible nonspacelike geodesic in (M,g) contains a pair of conjugate points.*
(3) *There exists a future trapped or past trapped set S in (M,g).*

This result of Hawking and Penrose implies the following result which is more similar to Theorem 12.41.

Theorem 12.44. *Let (M,g) be a chronological space–time of dimension $n \geq 3$ which satisfies the generic condition and the timelike convergence condition. If (M,g) contains a trapped set, then (M,g) is nonspacelike incomplete.*

Recall that a closed trapped surface is a compact spacelike submanifold of dimension $n-2$ for which the trace of both null second fundamental forms L_1 and L_2 is either always positive or always negative (cf. Definition 12.34).

Proposition 12.45. *Let (M,g) be a strongly causal space–time of dimension $n \geq 3$ which satisfies the condition $\mathrm{Ric}(v,v) \geq 0$ for all null vectors $v \in TM$. If (M,g) contains a closed trapped surface H, then at least one of (1) or (2) is true:*

(1) *(M,g) contains a trapped set.*
(2) *(M,g) is null incomplete.*

Proof. Assume that (M,g) is null complete. Corollary 12.36 then implies that one of $E^+(H)$ or $E^-(H)$ is compact. We consider the case that $E^+(H)$ is compact and define S by $S = E^+(H) \cap H$. We will show that S is a future trapped set. Notice that S is achronal since $E^+(H)$ is achronal and S is compact as the intersection of two compact sets.

Since $E^+(H) = J^+(H) - I^+(H)$, the set S will be nonempty if and only if H contains some points which are not in $I^+(H)$. But if H were contained in $I^+(H)$, there would be a finite cover of the compact set H by open sets $I^+(p_1), I^+(p_2), \ldots, I^+(p_n)$ with all $p_i \in H$. However, this would imply the existence of a closed timelike curve in (M,g) (cf. the proof of Proposition 3.10) which would contradict the strong causality of (M,g). Hence $S \neq \emptyset$.

In order to show that S is a future trapped set, it is sufficient to prove that $E^+(S) = E^+(H)$. We will demonstrate this by showing that $I^+(S) = I^+(H)$ and $J^+(S) = J^+(H)$. To this end, cover the compact set H with a finite number of open sets $U_1, U_2, \ldots, U_k$ of (M,g) such that each U_i is a convex normal neighborhood and no nonspacelike curve which leaves U_i ever returns.

Since H is a spacelike submanifold, we may also assume that each $U_i \cap H$ is achronal by choosing the U_i sufficiently small. Clearly, $I^+(S) \subseteq I^+(H)$ since $S \subseteq H$. To show $I^+(H) \subseteq I^+(S)$, assume that $q \in I^+(H) - I^+(S)$. Then there exists $p_1 \in H$ with $p_1 \ll q$. Now $p_1 \in U_{i(1)} \cap H$ for some index $i(1)$. Since $q \notin I^+(S)$, we have $p_1 \notin S$ and hence $p_1 \notin E^+(H)$. Thus there exists $p_2 \in H$ with $p_2 \ll p_1$. Since $U_{i(1)} \cap H$ is achronal, $p_2 \notin U_{i(1)}$. Now $p_2 \in U_{i(2)} \cap H$ for some $i(2) \neq i(1)$. Again $q \notin I^+(S)$ yields $p_2 \notin E^+(H)$. Thus there exists $p_3 \in H$ with $p_3 \ll p_2$. Furthermore, by construction of the sets U_i we have $p_3 \notin U_{i(1)} \cup U_{i(2)}$. Thus $p_3 \in U_{i(3)} \cap H$ for some $i(3)$ different from $i(1)$ and $i(2)$. Continuing in this fashion, we obtain an infinite sequence $p_1, p_2, p_3, \ldots$ in H with corresponding sets $U_{i(1)}, U_{i(2)}, U_{i(3)}, \ldots$ such that $i(j_1) \neq i(j_2)$ if $j_1 \neq j_2$. This contradicts the finiteness of the number of sets U_i of the given cover. Hence $I^+(S) = I^+(H)$. It remains to show that $J^+(S) = J^+(H)$. First $J^+(S) \subseteq J^+(H)$ as $S \subseteq H$. Thus assume that $q \in J^+(H) - J^+(S)$. Then $q \notin I^+(S) = I^+(H)$, and hence there is a future directed null curve from some point $p \in H$ to the point q. Since $p \in H$ and $p \notin I^+(H)$ as $p \leqslant q$ and

$q \notin I^+(H)$, we have $p \in E^+(H)$. Thus $p \in E^+(H) \cap H = S$ which yields $q \in J^+(S)$, in contradiction. This shows that S is a future trapped set.

A similar argument shows that if (M, g) is null complete and if $E^-(H)$ is compact, then $S = E^-(H) \cap H$ is a past trapped set. Thus the proposition is established. $\square$

It is possible for a trapped set to consist of just a single point. One way this may arise is as follows [cf. Hawking and Penrose (1970, p. 543), Hawking and Ellis (1973, pp. 266–267)]. Let (M, g) be a space–time of dimension $n \geq 3$ which satisfies the curvature condition $\text{Ric}(v, v) \geq 0$ for all null vectors $v \in TM$. Suppose that there exists a point p such that on every future directed null geodesic $\beta : [0, a) \to (M, g)$ with $\beta(0) = p$, the expansion $\overline{\theta}$ of the Lagrange tensor field $\overline{A}$ on $G(\beta)$ with $\overline{A}(0) = 0$ and $\overline{A}'(0) = E$ becomes negative for some $t_1 > 0$. Intuitively, each future directed null geodesic emanating from p has a point $\beta(t_1)$ in the future of p at which all future directed null geodesics are reconverging. Thus provided that the given null geodesic β can be extended to the parameter value $t_1 - (n-2)/\overline{\theta}(t_1)$, β has a future null conjugate point to $t = t_1$ along β. Hence $\beta(t) \in I^+(p)$ for all $t > t_1 - (n-2)/\overline{\theta}(t_1)$. Since the set of null directions at p is a compact set, it follows that $E^+(p) = J^+(p) - I^+(p)$ is compact provided that (M, g) is null geodesically complete. Thus we have obtained the following result.

Proposition 12.46. *Let (M, g) be a space–time of dimension $n \geq 3$ with* $\text{Ric}(v, v) \geq 0$ *for all null vectors $v \in TM$. Suppose that there exists a point $p \in M$ such that on every future directed null geodesic β from $p = \beta(0)$, the expansion $\overline{\theta}$ of the Lagrange tensor field $\overline{A}$ on $G(\beta)$ with $\overline{A}(0) = 0$ and $\overline{A}'(0) = E$ becomes negative for some $t_1 > 0$. Then at least one of* (1) *or* (2) *holds:*

(1) *$\{p\}$ is a trapped set, i.e., $E^+(p) = J^+(p) - I^+(p)$ is compact;*

(2) *(M, g) is null incomplete.*

We now consider the case of a compact connected spacelike hypersurface S without boundary in a space–time (M, g). If S is achronal, then $E^+(S) = S$ and S is a trapped set. On the other hand, it may happen that S is not achronal. In fact, an example may easily be constructed of a compact spacelike hypersurface S with $S \subseteq I^+(S)$ and hence $E^+(S) = \emptyset$ (cf. Figure 12.5).

A compact spacelike hypersurface S which is not achronal may be used to construct an achronal compact spacelike hypersurface $\widetilde{S}$ in a covering manifold $\widetilde{M}$ of M [cf. Geroch (1970), Hawking (1967, p. 194), O'Neill (1983)]. Thus if a space–time has a compact spacelike hypersurface, then there is a covering manifold of the given space–time which has a trapped set. But in proving the nonspacelike incompleteness of (M, g), we may work with covering manifolds just as well as with (M, g) since (M, g) is nonspacelike incomplete if and only if each covering manifold of (M, g) equipped with the pullback metric is nonspacelike incomplete.

These observations on covering spaces together with Theorem 12.44, Proposition 12.45, and Proposition 12.46 yield the following theorem [cf. Hawking and Penrose (1970, p. 544)].

Theorem 12.47. *Let (M, g) be a chronological space–time of dimension $n \geq 3$ which satisfies the generic condition and the timelike convergence condition. Then the space–time (M, g) is nonspacelike incomplete if any of the following three conditions is satisfied:*

(1) *(M, g) has a closed trapped surface.*

(2) *(M, g) has a point p such that each null geodesic starting at p is reconverging somewhere in the future (or past) of p.*

(3) *(M, g) has a compact spacelike hypersurface.*

Remark 12.48. Conditions (1) and (2) of Theorem 12.47 are reasonable cosmological assumptions. Robertson–Walker space–times with physically reasonable energy momentum tensors, positive energy density, and $\Lambda = 0$ have closed trapped surfaces [cf. Hawking and Ellis (1973, p. 353)] and thus satisfy (1). There is some astronomical evidence for (2) [cf. Hawking and Ellis (1973, p. 355)].

12.5 Smooth Boundaries

In this section we consider the relationship between causal disconnection, nonspacelike geodesic incompleteness, and points of the causal boundary $\partial_c M$

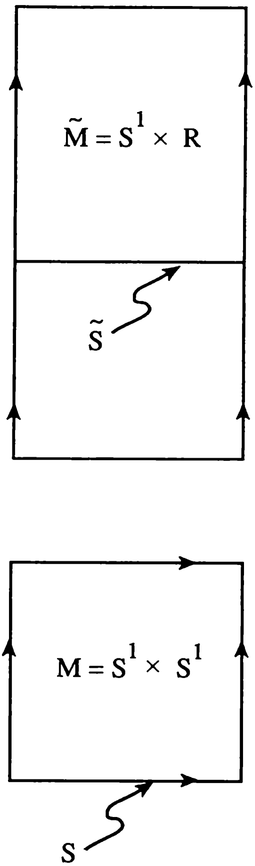


FIGURE 12.5. Let $M = S^1 \times S^1$ be given the Lorentzian metric $ds^2 = d\theta_1{}^2 - d\theta_2{}^2$. The set $S = \{(\theta_1, 0) : \theta_1 \in S^1\}$ is a compact spacelike submanifold of codimension one such that $E^+(S) = \emptyset$. Here S is not a trapped set because it is not achronal. However, (M, g) has a covering manifold $(\widetilde{M}, \widetilde{g})$ which contains a trapped set $\widetilde{S}$ which is diffeomorphic to S.

(cf. Section 6.4) at which the boundary is differentiable. Many of the more important space–times studied in general relativity have causal boundaries which are differentiable at a large number of points. For example, the differentiable part of the causal boundary of Minkowski space–time consists of $\mathfrak{I}^+$ and $\mathfrak{I}^-$ (cf. Figure 5.4). Since these sets correspond to null hypersurfaces, it is thus natural to call the points of $\mathfrak{I}^+$ and $\mathfrak{I}^-$ null boundary points. Penrose (1968) has used conformal methods to study smooth boundary points of Minkowski space–time and other space–times.

In general, we consider a space–time (M, g) with causal boundary $\partial_c M$ and let M^* denote the causal completion $M \cup \partial_c M$ of (M, g). Placing various causality conditions on (M, g) (e.g., stable causality) ensures that this completion may be given a Hausdorff topology such that the original topology on M agrees with the topology induced on M as a subset of M^* [cf. Hawking and Ellis (1973, pp. 220–221), Rube (1988), Szabados (1988), or Section 6.4].

Assume that $\overline{p} \in \partial_c M$ and let $U^* = U^*(\overline{p})$ be a neighborhood of $\overline{p}$ in M^*. Denote by (U, g) the metric g restricted to the set

$$U = U^* \cap M.$$

A *conformal representation* of $U^*(\overline{p})$ will be a space–time (M', g') and a homeomorphic embedding $f : U^* \to M'$ such that:

(1) $f \mid U$ is a smooth map; and

(2) There is a smooth function $\Omega : U \to \mathbb{R}$ such that $\Omega > 0$ and $\Omega g = f^* g'$ on U (Figure 12.6).

If the conformal representation $f : U^* \to M'$ maps U^* to a smooth manifold with boundary, then we will say that $\overline{p}$ is a *smooth boundary point*.

Definition 12.49. (*Smooth Causal Boundary Point*) Let $U^*(\overline{p})$ have a smooth conformal representation $f : U^* \to M'$ such that $f(U)$ is a smooth manifold with a smooth boundary $\partial(f(U))$ in M'. Then the point $\overline{p}$ is said to be a *smooth spacelike* (respectively, *null*, *timelike*) *boundary point* if the corresponding boundary $\partial(f(U))$ is a spacelike (respectively, null, timelike) hypersurface in (M', g').

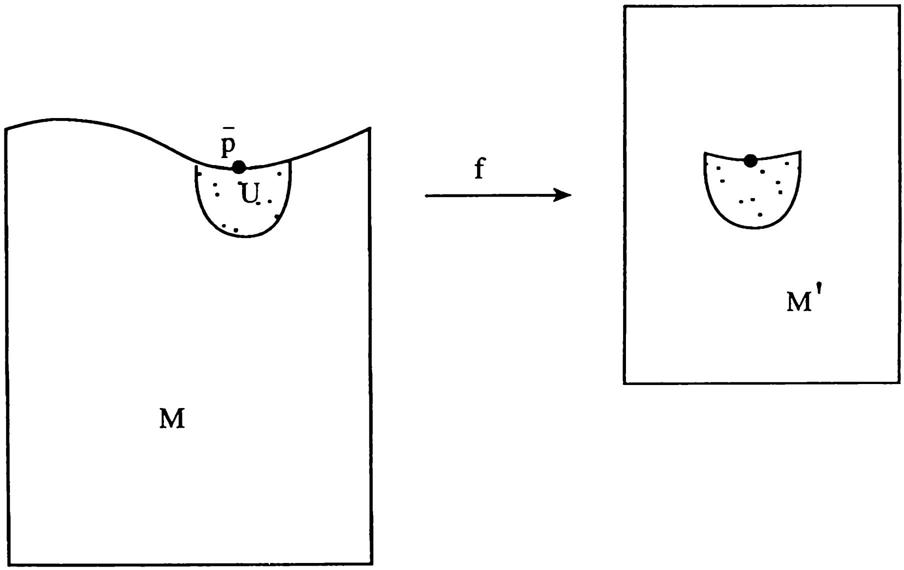


FIGURE 12.6. The space–time (M, g) has a causal boundary point $\overline{p}$, and $U^*(\overline{p})$ is a neighborhood of $\overline{p}$ in the causal completion $M^* = M \cup \partial_c M$ of (M, g). The homeomorphic embedding $f : U^* \to M'$ is a smooth conformal map on $U = U^*(\overline{p}) \cap M$.

If $\gamma : [a, b) \to U$ is a curve in M such that $\gamma(t) \to \overline{p} \in M^* - U$ as $t \to b$, then the curve γ is said to have the boundary point $\overline{p} \in M^*$ as an endpoint. Also if $\overline{p}$ is a smooth spacelike boundary point, then it is not hard to find a compact set K in M such that all inextendible nonspacelike curves with one endpoint at $\overline{p}$ must intersect K. In fact, K may be chosen as a compact, achronal spacelike hypersurface with boundary (cf. Figures 12.7 and 12.8). Furthermore, given any neighborhood $U^*(\overline{p})$ of $\overline{p}$ in M^*, the compact set K may be chosen to lie in $U = U^*(\overline{p}) \cap M$.

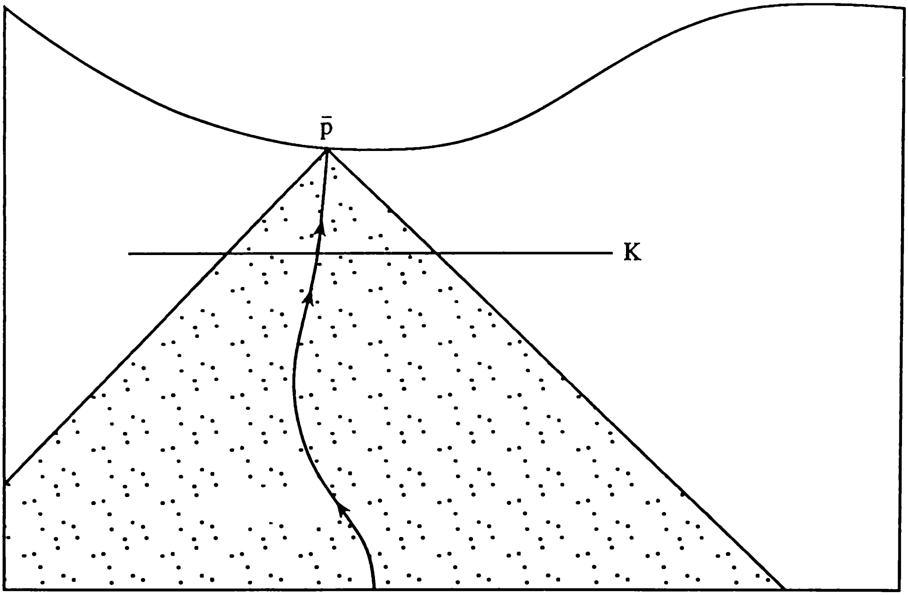


FIGURE 12.7. A space–time with a smooth spacelike boundary point $\overline{p}$ is shown. The compact set K is chosen such that any nonspacelike curve γ with one endpoint at $\overline{p}$ must intersect K.

Lemma 12.50. *If (M, g) is a space–time with a smooth spacelike boundary point, then (M, g) is causally disconnected.*

Proof. Let $\overline{p}$ be a smooth spacelike boundary point of (M, g). Choose a compact achronal set K such that any inextendible nonspacelike curve which has $\overline{p}$ as one endpoint must meet K. Let $\gamma : (-\infty, \infty) \to M$ be an inextendible nonspacelike curve with $\overline{p}$ as one endpoint and define $p_n = \gamma(-n)$, $q_n = \gamma(n)$ for each n. For all large n, the points p_n and q_n are causally disconnected by K (cf. the proof of Proposition 12.42). □

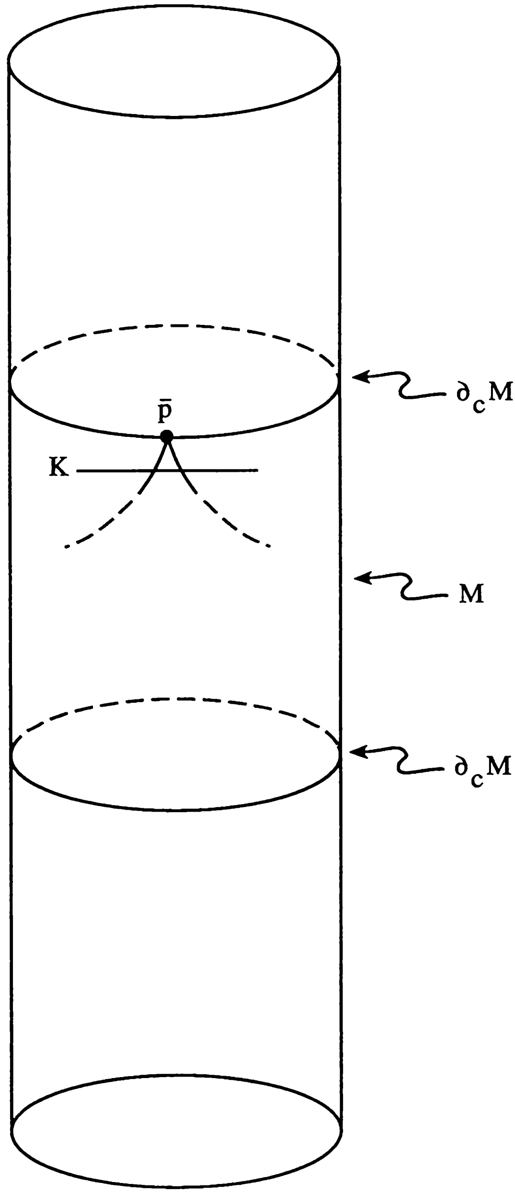


FIGURE 12.8. A closed Robertson–Walker cosmological model conformally represented as a subset of the Einstein static universe is shown. The set K is a causally disconnecting set.

Theorem 12.51. *Let (M, g) be a chronological space–time of dimension $n \geq 3$ which satisfies the generic condition and the timelike convergence condition. If (M, g) has a smooth spacelike boundary point, then (M, g) is nonspacelike incomplete.*

Proof. This follows from Lemma 12.50 and Theorem 12.41. □

Notice that if (M, g) has one smooth spacelike (respectively, null, timelike) boundary point, then (M, g) has uncountably many smooth spacelike (respectively, null, timelike) boundary points. For if $\overline{p} \in \partial_c M$ corresponds to the point $x \in \partial(f(U))$ under the given conformal representation $f : U^* \to M'$ of Definition 12.49, then points $y \in \partial(f(U))$ close to x will also represent smooth spacelike (respectively, null, timelike) boundary points in $\partial_c M$ under the given conformal representation and $\partial(f(U))$ has dimension $n - 1$ in M'. Hence M contains uncountably many smooth spacelike boundary points. Thus using the fact that a causally disconnecting set K may be chosen arbitrarily close to any smooth spacelike boundary point $\overline{p}$ and the result that a limit of maximal curves is maximal in a strongly causal space–time, we may also establish the following result.

Theorem 12.52. *Let (M, g) be a strongly causal space–time of dimension $n \geq 3$. Assume that (M, g) has a smooth spacelike boundary point and satisfies the generic condition and the timelike convergence condition. Then for each smooth spacelike boundary point $\overline{p}$, there is an incomplete nonspacelike geodesic which is inextendible and has $\overline{p}$ as an endpoint. Thus (M, g) has an uncountable number of incomplete, inextendible, nonspacelike geodesics.*

Proof. First, it follows from the preceding remarks that (M, g) contains an incomplete nonspacelike geodesic for each smooth nonspacelike boundary point. But we have just noted that if (M, g) contains one spacelike boundary point, it contains uncountably many spacelike boundary points. □

CHAPTER 13

GRAVITATIONAL PLANE WAVE SPACE–TIMES

It is well known and already mentioned in earlier chapters that for general space–times, geodesic connectedness and geodesic completeness are not well related. Indeed, in Figure 6.1 we have given an example of an elementary space–time which is geodesically complete yet have indicated in this figure a point q in $I^+(p)$ which is not joined to p by any geodesic whatsoever, let alone by a maximal timelike geodesic. More dramatically, in this same example a whole open neighborhood U about q exists, with U contained in $I^+(p)$, yet there is no geodesic from p to any point of U. Hence, despite the fact that this space–time is geodesically complete, the exponential map from p fails to map onto open subsets of $I^+(p)$. In Section 11.3, geodesic connectedness was explored from a general viewpoint and related to geodesic pseudoconvexity and geodesic disprisonment.

In this chapter we shall instead consider from the viewpoint of this book, and in particular as an illustration of the nonspacelike cut locus defined in Chapter 9, a class of space–times which originated as astrophysical examples yet which display a novel, very non-Riemannian geodesic behavior. Penrose (1965a) was apparently sufficiently impressed with the fact that the focusing behavior of future null geodesics issuing from appropriately chosen points p precluded this class of space–times from being globally hyperbolic that he chose the title "A remarkable property of plane waves in general relativity" for (1965a). In a series of articles by Ehrlich and Emch (1992a,b 1993), a detailed investigation of the behavior of all geodesics issuing from appropriately chosen p was conducted, and in so doing it was possible to place this class of space–times precisely in the standard causality ladder as being causally continuous yet not causally simple. Also, the future nonspacelike cut locus was determined and seen to coincide with the first future nonspacelike conjugate locus of p. Unfortunately, since this class of space–times was known in advance to fail to be globally hyperbolic, the theorems of Chapter 9 were not applicable here, only the basic concepts and definitions.

In Chapters 35 and 36 of Misner, Thorne, and Wheeler (1973), extensive discussions of gravitational waves as small scale ripples in the shape of space–time rolling across space–time is given by analogy with water waves as small ripples in the shape of the ocean's surface rolling across the ocean. These gravitational ripples are supposed to be produced by sources like binary stars, supernovae, or the gravitational collapse of a star to form a black hole. The exact plane wave solutions then arise as simplified models for this phenomenon, compromising between reality and complexity to paraphrase Misner, Thorne, and Wheeler (1973). We will refer the reader to this source rather than discussing the physics of these models in this chapter.

13.1 The Metric, Geodesics, and Curvature

This class of space–times, in the form in which we will describe them below, seems to originate with H. Brinkmann (1925) and stems from the work of Einstein [cf. Einstein and Rosen (1937)]. These metrics were brought to the renewed attention of the general relativity community in the late 1950's by I. Robinson [cf. Bondi, Pirani, and Robinson (1959) for a discussion of this history]. Let $M = \mathbb{R}^4$ with global coordinates (y, z, v, u). Heuristically, we may think of starting with the usual Minkowski coordinates and then setting

$$u = \frac{1}{\sqrt{2}}(t - x) \qquad \text{and} \qquad v = -\frac{1}{\sqrt{2}}(t + x). \tag{13.1}$$

In these coordinates, the Minkowski metric η has the form

$$\eta = 2du\,dv + dy^2 + dz^2. \tag{13.2}$$

We also make a choice of time orientation which will be convenient in the sequel, so that $\partial/\partial v$ is *past* directed null at all points. Now the class of "plane fronted waves" may be defined to be those metrics for $M = \mathbb{R}^4$ with global coordinates (y, z, v, u) which may be written in the form

$$g = \eta + H(y, z, u)du^2 \tag{13.3}$$

where $H(y, z, u)$ has an appropriate degree of regularity and is nonvanishing somewhere. The global coordinate $u : M \to \mathbb{R}$ plays an important role in the differential geometry of this class of space–times and is even more important for the subclass of gravitational plane waves introduced below in (13.9). The plane fronted metrics (13.3) have the interesting property that all have vanishing scalar curvature but need not be Ricci flat unless

$$\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) H(y, z, u) = 0. \tag{13.4}$$

No general statement can be made about the geodesic completeness of this class of space–times apart from the geodesics lying in the null hyperplanes

$$P(u_0) = \{(y, z, v, u) \in \mathbb{R}^4 : u = u_0\}. \tag{13.5}$$

On these hyperplanes, the ambient metric (13.3) is degenerate, so these are *not* "Lorentzian submanifolds" in the sense of O'Neill (1983) or of our previous Definition 3.47. Denoting y, z, v, u by $1, 2, 3, 4$ respectively, the only non-vanishing Christoffel symbols are

$$\Gamma^3_{44} = \frac{1}{2}\frac{\partial H}{\partial u}, \qquad \Gamma^3_{14} = \Gamma^3_{41} = -\Gamma^1_{44} = \frac{1}{2}\frac{\partial H}{\partial y},$$

and

$$\Gamma^3_{24} = \Gamma^3_{42} = -\Gamma^2_{44} = \frac{1}{2}\frac{\partial H}{\partial z}.$$

By direct calculation, it may be verified that any geodesic $c : (a, b) \to M$ satisfies $(x_4 \circ c)''(t) = 0$. Then in the case that $x_4 \circ c(t) = u_0$ for all t, one obtains that the geodesics in the null hyperplanes $P(u_0)$ are precisely straight lines lying in $P(u_0)$ of the form

$$c(s) = (a_1 s + a_2, b_1 s + b_2, c_1 s + c_2, u_0)$$

which are manifestly complete.

These geodesics are either spacelike or null; we will reserve the notation

$$\eta(s) = \eta_{y_0, z_0, u_0}(s) = (y_0, z_0, s, u_0) \tag{13.6}$$

for the member of the class of future directed null geodesics lying in $P(u_0)$ which passes through the point $(y_0, z_0, 0, u_0)$. An aspect of the differential geometry of plane fronted waves related to this class of null geodesics is the fact that the vector field $\operatorname{grad} u = \nabla u = \partial/\partial v$ is always a global parallel vector field, independent of the particular form of $H(y, z, u)$.

Already, it may be seen that the global coordinate $u : M \to \mathbb{R}$ has many aspects of a time function. First, we set things up in analogy with (13.1) so that $\partial/\partial v$ is past directed null. Thus it follows that u is strictly increasing along any smooth future directed timelike curve $c(t)$, for

$$(13.7) \qquad \partial/\partial t\,(u \circ c) = g(\nabla u, c'(t)) = g(\partial/\partial v, c'(t)) > 0$$

as $c'(t)$ is future timelike and $\partial/\partial v$ is past directed nonspacelike. Hence, any plane fronted wave is chronological. Furthermore, this coordinate function u is strictly increasing along any smooth future causal geodesic except for those null geodesics $\eta(s)$ of the form (13.6). For writing an arbitrary geodesic as $c(s) = (y(s), z(s), v(s), u(s))$, the geodesic differential equation reduces as noted above to $u''(s) = 0$ for this last component. Hence, if u is nonconstant, it may be rescaled to be $u(s) = s$, and

$$(13.8) \qquad c(s) = (y(s), z(s), v(s), s).$$

Thus u is indeed strictly increasing along all smooth null geodesics except for those of the form (13.6) lying in one of the null planes $P(u_0)$. Recall that a chronological space–time which fails to be causal contains a null geodesic $\beta : [0, 1] \to M$ with $\beta(0) = \beta(1)$. But such a segment cannot lie in one of the planes $P(u_0)$ because all null geodesics in these planes are injective by direct solution of the geodesic P.D.E.; hence β must have the form (13.8), but then the form of the last coordinate precludes $\beta(0) = \beta(1)$ from occurring. Thus the nature of the global coordinate u and the form of the geodesics in the null planes $P(u_0)$ ensure that all members of the class of plane fronted waves are causal space–times.

In Eardley, Isenberg, Marsden, and Moncrief (1986), it was shown that a space–time which is a non-trivial vacuum solution of the Einstein equations and which admits at least one non-homothetic conformal Killing vector field must be a plane fronted wave.

Definition 13.1. (*Gravitational Plane Wave*) The manifold $M = \mathbb{R}^4$ together with a plane fronted metric $g = \eta + H(u, y, z)du^2$ for which H takes the quadratic form

$$H(y, z, u) = f(u)(y^2 - z^2) + 2g(u)yz, \tag{13.9}$$

or in matrix notation

$$H(y, z, u) = \begin{pmatrix} y & z \end{pmatrix} \begin{pmatrix} f(u) & g(u) \\ g(u) & -f(u) \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} \tag{13.10}$$

with $f(u)$ or $g(u)$ not vanishing identically, is said to be a *gravitational plane wave*. The wave may be said to be *polarized* if $g(u) = 0$. The wave is commonly termed a *sandwich wave* if both functions f and g have compact support in u.

It is interesting that as a consequence of the quadratic nature of (13.9), or equivalently (13.10), the same linear second order system of differential equations

$$\begin{pmatrix} y \\ z \end{pmatrix}'' = \begin{pmatrix} f(u) & g(u) \\ g(u) & -f(u) \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} \tag{13.11}$$

arises in studying the Killing fields, the Jacobi fields, and the geodesics of these space–times. C. Ahlbrandt has commented that this is to be expected for the last two of these objects by the classical calculus of variations, given the quadratic nature of the metric in (13.9) and hence of the associated arc length functional. Also, all gravitational plane waves are geodesically complete and satisfy Ric $= 0$ but are not flat since either f or g is assumed to be nonvanishing. Thus, any geodesic conjugacy which arises may be attributed to the shear term rather than the Ricci curvature in equation (12.2) of Section 12.1. It may be checked directly that any future nonspacelike geodesic, other than one of the complete null geodesics of the form (13.6), experiences some nonzero timelike sectional curvatures and hence satisfies the generic condition (2.40).

We now turn to a more careful discussion of the geodesics of a plane wave space–time which are transverse to the null planes $P(u_0)$. As previously noted, such geodesics may be parametrized with $u(s) = s$, or equivalently, as

$$c(u) = (y(u), z(u), v(u), u). \tag{13.12}$$

Now in the case of the gravitational wave, the Christoffel symbols simplify to

$$\Gamma^3_{44} = \frac{1}{2}\left[f'(u)(y^2 - z^2) + 2g'(u)yz\right],$$
$$\Gamma^3_{14} = \Gamma^3_{41} = -\Gamma^1_{44} = f(u)y + g(u)z,$$
$$\text{and} \quad \Gamma^3_{24} = \Gamma^3_{42} = -\Gamma^2_{44} = g(u)y - f(u)z.$$

Hence, the y and z components of the geodesic P.D.E. read

$$y''(u) = f(u)y + g(u)z \quad \text{and} \tag{13.13}$$

$$z''(u) = g(u)y - f(u)z \tag{13.14}$$

as claimed above. Hence, it is immediate that if f and g are continuous, then differentiable solutions to (13.13) and (13.14) exist for all values of the parameter u. Rather than using the geodesic differential equation to solve for $v(u)$, it is convenient to give the following indirect argument. Suppose we wish to find the geodesic $c(t)$ with $g(c', c') = \lambda$. As $c'(u) = (y'(u), z'(u), v'(u), 1)$, this condition is expressed as

$$\lambda = g(c', c') = (y')^2 + (z')^2 + 2v'(u)\cdot 1 + [f(u)(y^2 - z^2) + 2g(u)yz]\cdot 1 \cdot 1$$

or

$$\lambda = (y')^2 + (z')^2 + (f(u)y + g(u)z)\, y + (g(u)y - f(u)z)\, z + 2v'.$$

Using (13.13) and (13.14), this last equation becomes

$$\lambda = (y')^2 + (z')^2 + y''y + z''z + 2v'$$

so that $v'(u) = \frac{1}{2}\left[-(yy')' - (zz')' + \lambda\right]$. Integration produces the equation

$$v(u) = \frac{1}{2}\left[-y(u)y'(u) - z(u)z'(u) + \lambda u + c\right] \tag{13.15}$$

which shows that since $y(u)$ and $z(u)$ are differentiable for all values of u, $v(u)$ may be calculated for all values of u by employing (13.15). This calculation, together with our previous knowledge about the geodesics of the null hyperplanes $P(u_0)$, gives an elementary direct proof of the geodesic completeness for this class of metrics; an alternate proof may be given based on a knowledge of the isometry group of this class of space–times. Equations (13.12)–(13.14) reveal that an important first step in understanding the geodesic behavior of these space–times is studying the system (13.11).

Thorough studies of the Killing vector fields and related isometry groups of these space–times have been given in Ehlers and Kundt (1962), Kramer, Stephani, MacCallum, Herlt, and Schmutzer (1980), and Ehrlich and Emch (1992a), among others. From the nature of the isometry group, it follows that if the behavior of the exponential map at $P_0 = (0,0,0,u_0)$ is understood, then as isometries map geodesics to geodesics, the same behavior obtains at an arbitrary point P with $u(P) = u_0$. With choice of initial conditions $y(u_0) = z(u_0) = v(u_0) = 0$, equation (13.15) simplifies to

$$v(u) = \tfrac{1}{2}\left[-y(u)y'(u) - z(u)z'(u) + \lambda(u - u_0)\right]. \tag{13.16}$$

In Emch and Ehrlich (1992a,b), a conjugacy index associated to this system with the preceding initial condition was discussed. For $u_j \in \mathbb{R}$, let

$$L(u_j) = \{\text{solutions } (y(u), z(u)) \text{ to (13.11) with } y(u_j) = z(u_j) = 0\}. \tag{13.17}$$

Definition 13.2. (*Conjugacy Index*) For distinct real numbers $u_0 < u_1$, define the *conjugacy index* $I(u_0, u_1)$ of O.D.E. system (13.11) with respect to u_0 and u_1 to be

$$I(u_0, u_1) = \dim\{L(u_0) \cap L(u_1)\} \tag{13.18}$$

which, given our dimension restriction, is either 0, 1, or 2. The pair $\{u_0, u_1\}$ is said to be *conjugate* provided $I(u_0, u_1) > 0$ and *astigmatic conjugate* provided that $I(u_0, u_1) = 1$. Also, $\{u_0, u_1\}$ is said to be a *first conjugate pair* provided that $I(u_0, u_1) > 0$ and $I(u_0, u) = 0$ for all $u \in (u_0, u_1)$. Also, introduce the notation

$$\operatorname{Conn}(P_0, u) = \{Q \in P(u) : \text{ there is a geodesic from } P_0 \text{ to } Q\}. \tag{13.19}$$

It is a consequence of Propositions 12.16 and 12.17 that every gravitational plane wave space–time admits a first conjugate pair. This was established by direct considerations in a symplectic context in Lemma 2.4 of Ehrlich and Emch (1992a).

In Ehrlich and Emch (1992a) the following facts were established by studying the exponential map and geodesic equation associated to a first astigmatic conjugate pair $u_0 < u_1$ for an arbitrary plane wave: let $P_0 \in P(u_0)$ be arbitrary. Then

(1) If $u_0 \leq u(Q) < u_1$, then there exists a unique geodesic (parametrized as discussed above) from P_0 to Q; hence $\dim \operatorname{Conn}(P_0, u) = 3$ for all u in $[u_0, u_1)$.
(2) $\dim \operatorname{Conn}(P_0, u_1) = 2$, and $\operatorname{Conn}(P_0, u_1)$ is in fact a plane in $P(u_1)$ every point of which is conjugate to P_0 by a one-parameter family of geodesics.
(3) If $u(Q) > u_1$, then $Q \in I^+(P_0)$.
(4) If $u_0 \leq u(Q) < u_1$, then $Q \in I^+(P_0)$ if and only if Q lies on a unique maximal timelike geodesic segment issuing from P_0.

Also, the first future nonspacelike conjugate locus and the future nonspacelike cut locus of P_0 in $\operatorname{Conn}(P_0, u_1)$ were seen to coincide. A nondifferentiable example was also given of a space–time of the above form with $\dim \operatorname{Conn}(P_0, u_1) = 1$, so this possibility was not ruled out in the considerations of Ehrlich and Emch (1992a). Now, the proofs of these results called upon the establishment of a symplectic framework somewhat outside the viewpoint of the subject matter of this book. Since it was seen in this reference that precisely the same behavior obtains in the polarized and nonpolarized cases, we will restrict our discussion in the next section to the polarized case for which the equations are somewhat less complicated and more easily dealt with in order to illustrate the above points.

13.2 Astigmatic Conjugacy and the Nonspacelike Cut Locus

In this section we turn to the study of astigmatic or index 1 conjugacy and the global geometric behavior of the geodesics issuing from P_0. As just mentioned, to avoid the complexities of dealing with the general case, which do not in any way affect the qualitative outcome of the geodesic geometry, we will restrict our attention to the somewhat simpler case of a polarized gravitational wave with $g(u) = 0$ and also assume $f(u) \geq 0$. Hence the metric takes the form

$$g = \eta + f(u)(y^2 - z^2)du^2, \tag{13.20}$$

and the geodesic differential equation for geodesics transverse to the null planes $P(u)$ uncouples as

$$y'' = f(u)y \quad \text{and} \tag{13.21}$$

$$z'' = -f(u)z. \tag{13.22}$$

Under the above assumption on f, standard O.D.E. theory shows that equation (13.22) admits a nontrivial first conjugate pair solution $z(u)$ with $z(u_0) = z(u_1) = 0$ and $z(u) > 0$ for u with $u_0 < u < u_1$. Fix such a number u_0. As $f \geq 0$, any nontrivial solution $y(u)$ to (13.21) with $y(u_0) = 0$ satisfies $y(u) \neq 0$ for all $u > u_0$.

Let us denote by $y_1(u)$ [respectively, $z_1(u)$] the solution to the O.D.E. (13.21) [respectively, (13.22)] with initial conditions

$$y_1(u_0) = z_1(u_0) = 0, \qquad \text{and} \qquad y_1'(u_0) = z_1'(u_0) = 1. \tag{13.23}$$

Since the isometry group of any gravitational plane wave space–time acts transitively on the null planes $P(u_0)$ for any $u_0 \in \mathbb{R}$, it is sufficient to understand the behavior of geodesics issuing from $P_0 = (0, 0, 0, u_0)$ to understand $\exp_P$ for any $P \in M$ with $u(P) = u_0$. Evidently, any geodesic $c(u)$ issuing from P_0 which is not a straight line lying in $P(u_0)$ and which satisfies $g(c', c') = \lambda$ may be described as

$$c(u) = \left(c_1 y_1(u), c_2 z_1(u), \frac{[-c_1{}^2 y_1(u) y_1{}'(u) - c_2{}^2 z_1(u) z_1{}'(u) + \lambda(u - u_0)]}{2}, u\right) \tag{13.24}$$

for some choice of constants c_1, c_2 in $\mathbb{R}$, in view of equation (13.16). For any u with $u_0 < u < u_1$, we have arranged that $y_1(u), z_1(u) > 0$. Hence, it is easy to check from (13.24) that not only does $\mathrm{Conn}(P_0, u) = P(u)$ for all such u, but also given any Q with $u(Q) \in [u_0, u_1)$, there is a *unique* geodesic from P_0 to Q. Now take any point $Q = (c, d, e, u_1)$ in $\mathrm{Conn}(P_0, u_1)$. Note that

$$c(u_1) = \left(c_1 y_1(u_1), 0, \frac{[-c_1{}^2 y_1(u_1) y_1{}'(u_1) + \lambda(u_1 - u_0)]}{2}, u_1 \right). \tag{13.25}$$

Hence, $d = 0$. Now c_1 is determined as $c_1 = c/y_1(u_1)$, possible since $y_1(u_1) > 0$. Given this determination of c_1, we then make $v(u_1) = e$ by choosing the constant λ to satisfy

$$\lambda = \frac{[2e + c^2 y_1{}'(u_1)/y_1(u_1)]}{[u_1 - u_0]} \tag{13.26}$$

and note that this setup is valid for any $c_2 \in \mathbb{R}$. Hence, we have not only established that

$$\mathrm{Conn}(P_0, u_1) = P(u_1) \cap \{z = 0\}, \tag{13.27}$$

but also letting c_2 vary over all possible values $-\infty < c_2 < +\infty$, we have shown that each point Q in $\mathrm{Conn}(P_0, u_1)$ is conjugate to P by this one-parameter family of geodesics as c_1 and $\lambda = g(c', c')$ are uniquely determined by Q, but c_2 is undetermined. The "unbounded" nature of this one-parameter family of geodesics precludes the geodesic system from being pseudoconvex, and also prevents this class of space–times from being globally hyperbolic [cf. Penrose (1965a)].

To recapitulate, we have found that as u increases from $u = u_0$, prior to reaching $P(u_1)$ each point Q with $u(Q) = u$ is joined to P_0 by a unique geodesic, but when $u = u_1$, the geodesic join set $\mathrm{Conn}(P_0, u_1)$ drops from three dimensions to two dimensions, and as that happens, every point in $\mathrm{Conn}(P_0, u_1)$ is conjugate to P_0 by a one-parameter family of geodesics as exhibited explicitly above.

The first future null conjugate locus of P_0 in the null hyperplane $P(u_1)$ may now be determined by taking $\lambda = 0$ in the above discussion. Noting that $y_1{}'(u_1)/y_1(u_1) > 0$ under our hypotheses, and denoting the y and v coordinates of

$$c(u_1) = \left(c_1 y_1(u_1), 0, \frac{[-c_1{}^2 y_1(u_1) y_1{}'(u_1)]}{2}, u_1 \right)$$

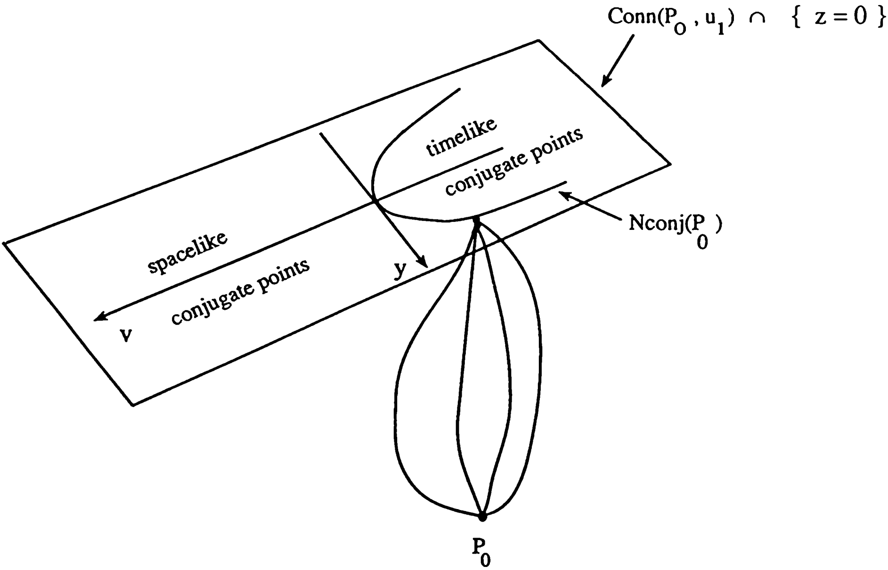


FIGURE 13.1. The conjugate locus of P_0 in $\text{Conn}(P_0, u_1) = P(u_0) \cap \{z = 0\}$ for a first conjugate pair $\{u_0, u_1\}$ and $P_0 = (0, 0, 0, u_0)$ is shown.

by Y and V respectively, we obtain the equation

$$V = -[(1/2)y_1'(u_1)/y_1(u_1)]Y^2, \qquad z = 0, \tag{13.28}$$

for the first future null conjugate locus of P_0 in $\text{Conn}(P_0, u_1)$. Hence, this null conjugate locus, $\text{Nconj}(P_0)$, is a planar parabola in $\text{Conn}(P_0, u_1)$. This same phenomenon is shown to occur for an arbitrary gravitational plane wave in Ehrlich and Emch (1992a, p. 214). The points in $\text{Conn}(P_0, u_1)$ which lie on timelike geodesics issuing from P_0 are precisely those which are in the same component of $\text{Conn}(P_0, u_1) - \text{Nconj}(P_0)$ as the focus of the parabola while those which lie on spacelike geodesics issuing from P_0 lie in the other component (cf. Figure 13.1).

More exactly, the following result is established for the general plane wave space–time in Ehrlich and Emch (1992a, p. 214).

Proposition 13.3. *In the presence of astigmatic conjugacy from P_0, the future null (respectively, nonspacelike) cut locus of P_0 is precisely the first future null (respectively, nonspacelike) conjugate locus to P_0 as described above. In particular, every point in the component of* $\mathrm{Conn}(P_0, u_1) - \mathrm{Nconj}(P_0)$ *which lies in the connected component containing the focus of the parabola* $\mathrm{Nconj}(P_0)$ *is a future timelike cut point to P_0.*

Proof. We first recall that these space–times have been known at least since Penrose (1965a) to fail to be globally hyperbolic; thus we may not use the existence of maximal nonspacelike geodesic segments in this proof. Instead, it is helpful to develop a version of the basic Proposition 3.4 which is more adapted to the sublevel sets of the global coordinate function $u : M \to \mathbb{R}$ rather than considering a small convex neighborhood centered at P_0. While u is *not* a time function in the sense of Section 3.2, u does possess the following two properties as considered in Ehrlich and Emch (1992a, pp. 200–204).

Definition 13.4. (*Quasi-Time Function*) A smooth function $f : N \to \mathbb{R}$ on an arbitrary space–time (N, g) is said to be a *quasi-time function* provided

(1) The gradient ∇f is everywhere nonzero, past directed nonspacelike, and
(2) Every null geodesic segment c such that $f \circ c$ is constant is injective.

From our description of the geodesics of gravitational plane wave space–times and the fact that $\nabla f = \partial/\partial v$ is past directed null, it is evident that the global coordinate function $u : M \to \mathbb{R}$ satisfies the properties of this definition. Property (1) implies that a quasi-time function is also a semi-time function as defined in Section 3.2, but it is the combination of both properties (1) and (2) in this definition which ensure that an arbitrary space–time which admits a quasi-time function must be causal. While a semi-time function prevents chronology from being violated, it does not, in general, prevent the existence of closed null geodesics.

Returning to the general setting of a plane wave, for any u_0, u_1 with $u_0 < u_1$, introduce the notation

$$S(u_0, u_1) = \{Q \in M : u_0 < u(Q) < u_1\}. \tag{13.29}$$

Arguing along the lines of (13.7), it may be seen that any such u-strip has the convexity property that if m, n are contained in $S(u_0, u_1)$, and $\gamma : [0,1] \to M$ is a future directed nonspacelike curve with $\gamma(0) = m$ and $\gamma(1) = n$, then

$$\gamma([0,1]) \subseteq S(u_0, u_1). \tag{13.30}$$

Now we restrict our attention to a strip $S(u_0, u_1)$ for which $\{u_0, u_1\}$ is a first conjugate pair. Then by our above considerations, $\exp_{P_0}$ is invertible on this particular strip. Thus the Synge world function $\Phi : S(u_0, u_1) \to \mathbb{R}$ given by

$$\Phi(m) = g\big(\exp_{P_0}^{-1}(m), \exp_{P_0}^{-1}(m)\big) \tag{13.31}$$

is differentiable on $S(u_0, u_1)$. Using the Synge world function, define the following partition of $S(u_0, u_1)$. Put

$$N = \Phi^{-1}(0) = \{\text{points in } S(u_0, u_1) \text{ that are joined to } P_0 \text{ by a null geodesic}\},$$

$$T = \{m \in S(u_0, u_1) : \Phi(m) < 0\} = \{\text{points in } S(u_0, u_1) \text{ which are joined to } P_0 \text{ by a timelike geodesic}\},$$

and

$$S = \{m \in S(u_0, u_1) : \Phi(m) > 0\} = \{\text{points in } S(u_0, u_1) \text{ which are joined to } P_0 \text{ by a spacelike geodesic}\}.$$

Even though T may not be contained in a convex normal neighborhood of P_0, we still have $\nabla\Phi$ existing and being future timelike on all of T. Now the promised technical lemma may be stated.

Sublemma 13.5. *Let $\gamma : [0,1] \to (M,g)$ be a piecewise smooth, future directed, nonspacelike curve with $\gamma(0) = P_0$, $\gamma'(0)$ future timelike, and with $\gamma(1)$ lying in $S(u_0, u_1)$. Then $\gamma((0,1]) \subseteq T$.*

Proof of Sublemma 13.5. First, by the u-causal convexity mentioned above, we have $\gamma((0,1]) \subseteq S(u_0, u_1)$. If we now take a small convex neighborhood U of P_0 and choose $t_0 > 0$ such that $\gamma([0,t_0]) \subseteq U$, then by the basic Proposition 3.4, we have $\Phi(\gamma(s)) < 0$ for any s with $0 < s \leq t_0$, and in particular, $\Phi(\gamma(t_0)) < 0$. But now $\Phi \circ \gamma$ must be decreasing on $[t_0, 1]$ since for any $t \geq t_0$, $\partial/\partial t(\Phi(\gamma(t))) = g(\gamma'(t), \nabla\phi|_{\gamma(t)}) < 0$ as $\gamma'(t)$ is future nonspacelike and $\nabla(\Phi)$ is future timelike. Hence $\gamma((0,1]) \subseteq T$. □

Corollary 13.6. *Suppose $R \in I^+(P_0) \cap S(u_0, u_1)$. Then there exists a unique maximal timelike geodesic from P_0 to R.*

Proof. Under the hypotheses on R, there exists a curve γ from P_0 to R satisfying the hypotheses of Sublemma 13.5. Hence $R \in T$, and the unique geodesic c in $S(u_0, u_1) \cup \{P_0\}$ from P_0 to R must be timelike. Now if $\sigma : [0,1] \to M$ is an arbitrary timelike curve with $\sigma(0) = P_0$ and $\sigma(1) = R$, then by the u-causal convexity, $\sigma((0,1])$ is contained in $S(u_0, u_1)$. Hence σ may be lifted to $T_{P_0}M$, and thus the usual Gauss Lemma type arguments imply that σ cannot be longer than the timelike geodesic c from P_0 to R. Hence, this geodesic must be maximal. □

Conclusion of the Proof of Proposition 13.3. No future nonspacelike geodesic $c(u) = (y(u), z(u), v(u), u)$ issuing from P_0 can be maximal past $u = u_1$ because c is conjugate to P_0 at $c(u_1)$ as noted above. Suppose then that $c : [u_0, u_1) \to \{P_0\} \cup S(u_0, u_1)$ is a future timelike geodesic with $c(u_0) = P_0$ which contains a timelike cut point $Q = c(u_2)$ to P_0 with $u_0 < u_2 < u_1$. Choose $\delta > 0$ so that $u_2 + \delta < u_1$, and let $R = c(u_2 + \delta)$. By the general theory of timelike cut points, $c\,|\,[u_0, u_2 + \delta]$ cannot be maximal. But this then contradicts Corollary 13.6 which guarantees that $c\,|\,[u_0, u_2 + \delta]$ is the unique maximal timelike geodesic from P_0 to R.

Now consider the case of a null geodesic $c : [u_0, u_1) \to \{P_0\} \cup S(u_0, u_1)$ issuing from P_0. Suppose $Q = c(u_2)$ is a null cut point along c to P_0 with

$u_0 < u_2 < u_1$. Choose $\delta > 0$ such that $u_2 + \delta < u_1$, and put $S = c(u_2 + \delta)$. By the general theory of null cut points, there exists a timelike curve from P_0 to S satisfying the hypotheses of Sublemma 13.5. Hence $S \in T$ by this sublemma. Thus P_0 is joined to S by a timelike geodesic as well as the given null geodesic c. But this contradicts the fact established above that every point in $S(u_0, u_1)$ is joined to P_0 by exactly one geodesic. □

To complete the discussion of the nonspacelike cut locus, it remains to consider the null geodesic $\eta(s) = (0, 0, s, u_0)$ issuing from P_0 and lying in the null plane $P(u_0)$ itself.

Remark 13.7. All the future null geodesics η defined as in (13.6) contained in one of the null hyperplanes $P(u_0)$ are globally maximal. For suppose such a null geodesic η is not globally maximal. Then η contains a pair of points $p = \eta(s)$, $q = \eta(t)$, with $s < t$, and a smooth, future timelike curve $\gamma : [0, 1] \to M$ with $\gamma(0) = p$ and $\gamma(1) = q$. But then $u(p) < u(q)$ as u is a quasi-time function and γ is future timelike, contradicting $u(p) = u(q) = u_0$.

13.3 Astigmatic Conjugacy and the Achronal Boundary

Now we will change our point of view slightly and investigate how astigmatic conjugacy relates to certain familiar characterizations of achronal boundaries in terms of null geodesic behavior. As in the last section, let $u_0 < u_1$ denote a first astigmatic conjugate pair for a polarized gravitational wave metric $g = \eta + f(u)(y^2 - z^2)du^2$, and let $P_0 = (0, 0, 0, u_0)$ and $Q_0 = (0, 0, 0, u_1)$.

We have briefly discussed the concept of future sets and achronal boundaries in Chapter 3, beginning with Definition 3.6. Some of the techniques discussed there will be helpful in this section. We now introduce a subset of $\text{Conn}(P_0, u_1)$ which has been termed the "null tail" in Ehrlich (1993).

Definition 13.8. (*Null Tail*) For any R with $u(R) = u_1$ and $R \in \text{Nconj}(P_0)$, let β_R denote the *past directed*, past complete, null geodesic ray in $\text{Conn}(P_0, u_1)$ issuing from R. Then the *null tail* $NT = NT(P_0)$ of P_0 in $\text{Conn}(P_0, u_1)$ (cf. Figure 13.2) is given by

$$NT = \bigcup_{R \in \text{Nconj}(P_0)} \beta_R. \tag{13.32}$$

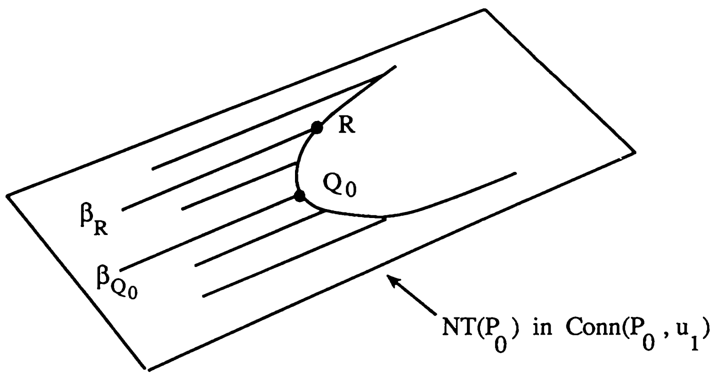


FIGURE 13.2. The null tail $NT(P_0)$ in $P(u_0)$ for a first astigmatic conjugate pair $u_0 < u_1$ is shown.

Points Q on the null tail $NT(P_0)$ then fall into one of two categories: first, those points lying on $\mathrm{Nconj}(P_0)$ and thus joined to P_0 by a one-parameter family of null geodesics; and second, those points Q not lying on $\mathrm{Nconj}(P_0)$ which are thus joined to P_0 by a one-parameter family of spacelike geodesics. Penrose (1965a) observed from the "noncompact" behavior of the family of null geodesics of the type (13.24) from P_0 which refocus at Q_0 that β_{Q_0} happens to be a limit curve of this family of null geodesics from P_0 to Q_0, and even more remarkably, this behavior prevents this class of space–times from being globally hyperbolic. In Ehrlich and Emch (1992a, p. 218), we noted that this observation may be used as a step in the proof of the following result.

Lemma 13.9. *Suppose* $u_0 < u_1$ *is a first astigmatic conjugate pair and* $u(Q) > u_1$. *Then* $Q \in I^+(P_0)$.

Proof. First, choose $c > 0$ such that $(u_1, u_1 + c]$ contains no u-value conjugate to $u = u_1$. Then from the explicit form of the geodesics for gravitational plane waves, it may be calculated that for any Q in the strip $S(u_1, u_1 + c)$, there exists a future timelike geodesic from a point R on the null ray $\beta_{Q_0} =$

$\{(0,0,s,u_1) : s \in \mathbb{R}\}$ to Q. But then since $R \in \overline{I^+(P_0)}$ as Penrose observed, we have $Q \in I^+(P_0)$ by future set theory, as $\overline{I^+(P_0)} = \{m \in M : I^+(m) \subseteq I^+(P_0)\}$ (cf. Proposition 3.7 or Corollary 3.8). Now given an arbitrary Q with $u(Q) > u_1$, simply take any past directed timelike geodesic $c(u)$ issuing from Q, and from knowledge of the quasi-time function u, we know for any u_2 in $(u_1, u_1 + c)$, that $S = c(u_2)$ satisfies $S \ll Q$ and $S \in S(u_1, u_1 + c) \subseteq I^+(P_0)$. Hence, $Q \in I^+(P_0)$ as well. □

Now $F = I^+(P_0)$ is a particular example of a future set, and the corresponding achronal boundary is given by

$$B = \partial(I^+(P_0)) = \overline{I^+(P_0)} - I^+(P_0). \tag{13.33}$$

The characterization [cf. Penrose (1972, Section 5)] of a general achronal boundary in terms of what relativists call the *null geodesic generators* translates in our simpler context of (13.33) as follows.

Theorem 13.10. *Let (N, g) be an arbitrary space–time, and let $B = \partial(I^+(P_0))$ for any $P_0 \in N$. Suppose $R \in B - \{P_0\}$. Then there exists a null geodesic β contained in B with future endpoint R which satisfies either*

(1) *$\beta : [0,1] \to N$ with $\beta(0) = P_0$ and $\beta(1) = R$, or*
(2) *$\beta : (-a, 1] \to B$ is past inextendible and $\beta(1) = R$.*

It is possible for a particular point on the achronal boundary B to possess geodesics of both types (1) and (2) above. The null tail $NT(P_0)$ defined above happens to be a subset of the achronal boundary $B = \partial(I^+(P_0))$, and points R of $NT(P_0)$ which lie on $\text{Nconj}(P_0)$ as well have null geodesics with endpoint R of both type (1) and type (2) of Theorem 13.10. On the other hand, points R of the null tail $NT(P_0)$ which do not lie on the null conjugate locus $\text{Nconj}(P_0)$ are future endpoints of a null geodesic β only of type (2).

For this simple example of a future set, note that points R which satisfy alternative (1) lie in the future horismos $E^+(P_0) = J^+(P_0) - I^+(P_0)$ of P_0. Also, it should be noted that if a space–time (N, g) is causally simple, so that $J^+(Q) = \overline{J^+(Q)} = \overline{I^+(Q)}$ for any Q in N, then any future set B defined as in (13.33) satisfies $B = \partial(I^+(Q)) = E^+(Q)$. Thus to get an example of such an achronal boundary B which contains points R satisfying alternative (2) but not alternative (1) of Theorem 13.10, it is necessary to consider space–times which fail to be causally simple. Fortunately, all gravitational plane waves are known to fail to be causally simple, essentially because of the behavior mentioned above that the null geodesic ray β_{Q_0}, apart from its terminal point Q_0, is in $\overline{J^+(P_0)}$ but not in $J^+(P_0)$.

At present, we know that $I^+(P_0)$ and $B = \partial(I^+(P_0))$ have the following properties for the first astigmatic conjugate pair $u_0 < u_1$ and $P_0 = (0,0,0,u_0)$.

(1) If $R \in I^+(P_0)$, then $u(R) > u_0$.
(2) If $u_0 \leq u(R) < u_1$, then $R \in I^+(P_0)$ if and only if R lies on a unique maximal timelike geodesic segment c with $c(u_0) = P_0$ and $c(u(R)) = R$.
(3) If $u(R) > u_1$, then $R \in I^+(P_0)$ automatically.
(4) If $u(R) = u_1$ and $R \in \text{Conn}(P_0, u_1)$, then $R \in I^+(P_0)$ if and only if R lies on a uniquely determined 1-parameter family of maximal timelike geodesic segments from P_0 to R.

What is not settled is the question of what other points R in $P(u_1)$ which do *not* lie on geodesics issuing from P_0, i.e., $z(R) \neq 0$, are contained in $I^+(P_0)$, or in simpler terms, what happens to the chronological future of P_0 as astigmatic conjugacy occurs when $u = u_1$?

The corresponding facts for the achronal boundary $B = \partial(I^+(P_0))$ are as follows.

(5) If $R \in B$, then $u_0 \leq u(R) \leq u_1$.
(6) If $u_0 \leq u(R) < u_1$, then $R \in B$ if and only if R is the endpoint of a unique maximal null geodesic segment from P_0 to R.
(7) For R in $\text{Conn}(P_0, u_1)$, it is known that $R \in B$ if and only if R is contained in the null tail $NT(P_0)$.

In the case of B, then, the issue is whether $B \cap P(u_1) = NT(P_0)$, or equivalently, whether $P(u_1) - \text{Conn}(P_0, u_1) \subseteq I^+(P_0)$.

It may be seen for a subclass of polarized gravitational waves, which were termed "unimodal" in Ehrlich and Emch (1992a) and Ehrlich (1993), that these related issues may be settled affirmatively.

Definition 13.11. (*Unimodal Gravitational Plane Wave Metric*) A polarized gravitational plane wave metric $g = \eta + f(u)(y^2 - z^2)du^2$ is said to be *unimodal* if $f(u) \geq 0$ and there exist $A > 0$ and $\overline{u} \in \mathbb{R}$ with the following properties:

$$A = \int_{-\infty}^{\overline{u}} [f(u)]^{1/2}\, du = \int_{\overline{u}}^{+\infty} [f(u)]^{1/2}\, du; \tag{13.34}$$

$$0 < A \leq \frac{\pi}{2}; \quad \text{and} \tag{13.35}$$

$$(u - \overline{u})f'(u) \leq 0 \qquad \text{for all } u \in \mathbb{R}. \tag{13.36}$$

Requirement (13.35) serves to ensure that in a certain Prüfer type transformation for the solution of the O.D.E. (13.22), a continuous determination of arctan may be made [cf. Ehrlich and Emch (1992a, p. 188)]. What we need for our purposes in this section, which is where the assumption of unimodality enters into the discussion below, is the following variant of a Sturm Separation Theorem, which is implicit in the proof of Theorem 2.14 of Ehrlich and Emch (1992a, p. 187).

Lemma 13.12. *Let $u_0 < u_1$ be a first astigmatic conjugate pair for a given unimodal gravitational wave. Then there exist $\delta > 0$ and a strictly increasing continuous function $a : [0, \delta) \to [0, +\infty)$ with $a(0) = 0$ and such that*

$$\{u_0 + \epsilon, u_1 + a(\epsilon)\} \tag{13.37}$$

is a first astigmatic conjugate pair for all ϵ in $[0, \delta)$.

Now we are ready for

Theorem 13.13. *Let $u_0 < u_1$ be a first astigmatic conjugate pair for a unimodular polarized gravitational wave. Let $P_0 = (y_0, z_0, v_0, u_0)$ be an arbitrary point of the null plane $P(u_0)$. If $R \in P(u_1)$ does not lie on any geodesic issuing from P_0, then $R \in I^+(P_0)$. Hence, the null tail $NT(P_0)$ of P_0 satisfies*

$$NT(P_0) = \partial(I^+(P_0)) \cap P(u_1). \tag{13.38}$$

Proof. From the nature of the isometry group, it suffices to make explicit calculations with $P_0 = (0,0,0,u_0)$ as above. With this particular choice of P_0, for any gravitational plane wave metric with $g(u) = 0$ we have that $\beta(s) = (0,0,0,s)$ is a future directed null geodesic issuing from P_0. Let P_ϵ denote the point given by $P_\epsilon = \beta(u_0 + \epsilon) = (0,0,0,u_0+\epsilon)$. Further, let $y_\epsilon(u)$ and $z_\epsilon(u)$ denote respectively the unique solutions to the IVP's

$$y'' - f(u)y = 0, \qquad y(u_0+\epsilon) = 0, \quad y'(u_0+\epsilon) = 1 \quad \text{and} \tag{13.39}$$

$$z'' + f(u)z = 0, \qquad z(u_0+\epsilon) = 0, \quad z'(u_0+\epsilon) = 1 \tag{13.40}$$

which arise in studying the exponential map from P_ϵ. In view of (13.24) translated to P_ϵ, if a point $Q = (y,z,v,u)$ with $u > u_0 + \epsilon$ lies on a geodesic $c(u)$ issuing from P_ϵ with $g(c',c') = \lambda$, then

$$2v = -\frac{y_\epsilon{}'(u)}{y_\epsilon(u)}\, y^2 - \frac{z_\epsilon{}'(u)}{z_\epsilon(u)}\, z^2 + \lambda(u - u_0 - \epsilon). \tag{13.41}$$

Now fix $R = (a,b,c,u_1)$ in $P(u_1)$ which does not lie on any geodesic issuing from P_0, ensuring $b \neq 0$. If we show that R lies on some future timelike geodesic issuing from P_ϵ for some $\epsilon > 0$, then R lies on a future null geodesic from P_0 to P_ϵ, followed by a future timelike geodesic from P_ϵ to R, whence $P_0 \ll R$ as desired.

But the point R will lie on a timelike geodesic from P_ϵ for some $\epsilon > 0$ provided that the inequality

$$2c < -\frac{y_\epsilon{}'(u_1)}{y_\epsilon(u_1)}\, a^2 - \frac{z_\epsilon{}'(u_1)}{z_\epsilon(u_1)}\, b^2 \tag{13.42}$$

is satisfied for some $\epsilon > 0$. Since $b^2 > 0$, this will ensue provided that

$$\frac{y_\epsilon{}'(u_1)}{y_\epsilon(u_1)} \tag{13.43}$$

is uniformly bounded as $\epsilon \to 0$ while

$$\lim_{\epsilon \to 0} \frac{z_\epsilon{}'(u_1)}{z_\epsilon(u_1)} = -\infty. \tag{13.44}$$

Choose a constant $B > 0$ so that $0 \leq f(u) \leq B^2$ for u with $u_0 - 1 \leq u \leq u_1 + 1$. Also as $f(u) \geq 0$, we have that $y_\epsilon(u)$ and $y_\epsilon{}'(u)/y_\epsilon(u)$ are both positive for any $u > u_0 + \epsilon$. Hence (13.43) follows by the usual Rauch comparison theory techniques.

The more technical part of the proof is to establish (13.44). Here there are two cases, both of which occur in explicit examples: first, $u_1 \notin \operatorname{supp}(f(u))$, and second, $u_1 \in \operatorname{supp}(f(u))$. In either case, let $a(\epsilon)$ be as in (13.37).

Consider first the case that $u_1 \notin \operatorname{supp}(f(u))$, and so also $u_1 + a(\epsilon) \notin \operatorname{supp}(f(u))$ for all $\epsilon > 0$ and $z''(u) = 0$ for all $u \geq u_1$. Hence there exist constants λ_ϵ so that

$$z_\epsilon(u) = \lambda_\epsilon\big(u - (u_1 + a(\epsilon))\big) \tag{13.45}$$

for all $u \geq u_1$. Consequently,

$$\frac{z_\epsilon{}'(u_1)}{z_\epsilon(u_1)} = \frac{-1}{a(\epsilon)}, \tag{13.46}$$

so that limit (13.44) is immediate.

Finally, suppose that $u_1 \in \operatorname{supp}(f(u))$. L. Flaminio has remarked that the analysis step (13.46) should remain valid because the Jacobi solution $z(u)$ near its zero at $u_1 + a(\epsilon)$ is known to be asymptotically linear. A detailed calculation demonstrating that this intuition is valid may be given by drawing on the variant of the Prüfer transform technique used as in the proof of Theorem 2.14 of Ehrlich and Emch (1992a). The IVP's (13.40) in the (u, z) plane are transformed into the (s, θ) plane as in Theorem 2.14 in order to carry out the details of the analysis. The crux of the matter is that the transform of $z_\epsilon(u)$ to $\theta_\epsilon(s)$ enables a uniform bound on $\theta_\epsilon{}'(s)$ to be given because of the presence of the term $\sin[\theta_\epsilon(s)]$ on the right hand side of equation (2.83) of Ehrlich and Emch (1992a). □

CHAPTER 14

THE SPLITTING PROBLEM IN GLOBAL LORENTZIAN GEOMETRY

An important aspect of global Riemannian geometry over the past 25 years has been the investigation of *rigidity* in connection with curvature inequalities. This concept was first widely disseminated in 1975 in the first two paragraphs of the preface to the influential monograph by J. Cheeger and D. Ebin (1975), *Comparison Theorems in Riemannian Geometry*:

> In this book we study complete riemannian manifolds by developing techniques for comparing the geometry of a general manifold M with that of a simply connected model space M_H of constant curvature H. A typical conclusion is that M retains particular *geometrical* properties of the model space under the assumption that its sectional curvature K_M, is bounded between suitable constants. Once this has been established, it is usually possible to conclude that M retains *topological* properties of M_H as well.
>
> The distinction between strict and weak bounds on K_M is important, since this may reflect the difference between the geometry of say the sphere and that of euclidean space. However, it is often the case that a conclusion which becomes false when one relaxes the condition of strict inequality to weak inequality can be shown to fail only under very special circumstances. Results of this nature, which are known as *rigidity theorems*, generally require a delicate global argument.

A first example is provided by the topological sphere theorem established by H. Rauch (1951), M. Berger (1960), and W. Klingenberg (1959, 1961, 1962), among other authors.

Topological Sphere Theorem. *Let (N, h) be a simply connected, complete Riemannian manifold of dimension $n \geq 2$. Suppose the sectional curvatures of (N, h) satisfy the curvature pinching condition*

$$d \cdot A < K(\sigma) \leq A \tag{14.1}$$

for all two-planes σ in $G_2(M)$, where d satisfies

$$d > \frac{1}{4} \tag{14.2}$$

and $A > 0$ is arbitrary. Then N is homeomorphic to the sphere S^n.

Notice that in this result we have a strict inequality in (14.2) and also in the left inequality of (14.1). Now the rigidity in this context was first proved by M. Berger under the weaker pinching hypothesis

$$\frac{1}{4} A \leq K(\sigma) \leq A. \tag{14.3}$$

Berger obtained that either the previous topological conclusion held and that N would still be homeomorphic to S^n, or if the previous conclusion *failed* to hold, and hence N was no longer homeomorphic to S^n, then N had not just to be homeomorphic but had to be *isometric* to a compact rank one Riemannian symmetric space.

Here is a second example more germane to our context. It follows from the work of Gromoll and Meyer (1969) that a complete Riemannian manifold of dimension $n \geq 2$ with everywhere positive Ricci curvatures

$$\operatorname{Ric}(v, v) > 0 \qquad \text{for all nonzero } v \text{ in } TM \tag{14.4}$$

is connected at infinity. Rigidity now enters into this situation in the following manner. Suppose only that $\operatorname{Ric}(v, v) \geq 0$ and that (N, h) *fails* to be connected at infinity. Then since (N, h) is assumed to be complete, there exists what is often termed a (geodesic) *line*, i.e., a complete geodesic $c : \mathbb{R} \to (N, h)$ joining any two different ends of N and minimizing distance between any two of its points. But then the Cheeger–Gromoll Splitting Theorem (1971) may be applied to the line c to establish that (N, h) is *isometric* to a product manifold.

Cheeger–Gromoll Splitting Theorem. *Let (N, h) be a complete Riemannian manifold of dimension $n \geq 2$ which satisfies the curvature condition*

$$\text{(14.5)} \qquad \operatorname{Ric}(v, v) \geq 0 \qquad \textit{for all } v \textit{ in } TM$$

and which contains a complete geodesic line. Then (N, h) may be written uniquely as an isometric product $N_1 \times \mathbb{R}^k$ where N_1 contains no lines and $\mathbb{R}^k$ is given the standard flat metric.

Thus in this example the curvature rigidity arises as condition (14.4) is weakened to condition (14.5).

In 1932 Busemann (1932) introduced an analytic method of studying generalizations of the classical horospheres of hyperbolic geometry to a very general class of metric geometries. In the proof of the Cheeger–Gromoll Splitting Theorem given in Cheeger and Gromoll (1971), this particular function was independently rediscovered during the course of the proof as a key ingredient. At about the same time, the Busemann function had also been seen to be a useful tool in the study of complete Riemannian manifolds of nonpositive sectional curvature (cf. Eberlein and O'Neill (1973), Eberlein (1973a), where the function was named the "Busemann function" in honor of its discoverer). This function has played a prominent role continuing up to the present time in global Riemannian geometry.

In Chapter 12 singularity theorems for space–times which satisfy certain global geometric conditions and certain curvature conditions have been discussed. A simple prototype may be stated as a reminder.

Prototype Singularity Theorem. *Let (M, g) be a space–time of dimension $n > 2$ which satisfies the following three conditions:*

1. *(M, g) contains a compact Cauchy surface;*
2. *(M, g) satisfies $\operatorname{Ric}(v, v) \geq 0$ for all timelike (or all nonspacelike) v; and*
3. *Every inextendible nonspacelike geodesic satisfies the generic condition (cf. Definition 12.7 and Theorem 12.18).*

Then (M, g) contains an incomplete nonspacelike geodesic.

Now in this result the "strict curvature inequality" is condition (3), which requires for each inextendible nonspacelike geodesic that a certain curvature quantity be *nonzero* at some point of the given geodesic. Thus, in this context, curvature rigidity would suggest studying space–times which satisfy all the conditions of the Prototype Singularity Theorem but condition (3).

Earlier in an essay on singularities in general relativity, Geroch (1970b, pp. 264–265) had suggested for inextendible closed universes satisfying the Einstein equations that, generically, such space–times should be timelike geodesically incomplete [cf. the middle column in Figure 2 on p. 266 of Geroch (1970b)]. Geroch's conjectured viewpoint was stated as

> Thus, we expect that the diagram for closed universes will be almost entirely black. There are, however, at least a few white points: there exist closed, geodesically complete flat spacetimes ... Perhaps there are a few other nonsingular closed universes, but these may be expected to appear either as isolated points or at least regions of lower dimensionality in an otherwise black diagram.

Here Geroch is representing diagrammatically the space of all solutions to Einstein's equations in the above discussion, with singular space–times represented as a black dot and non-singular, inextendible space–times represented as a white dot. Further, in Geroch (1970b, p. 288) the following problem was posed:

> Prove or find a counterexample: Every inextendible space–time containing a compact spacelike 3-surface and whose stress-energy tensor satisfies a suitable energy condition is either G-singular or flat. (The energy condition must be strong enough to exclude the Einstein universe. Perhaps one should first prove that the 3-surface cannot be, topologically, a 3-sphere.)

Galloway and Horta (1995) summarize Geroch's ideas here as "spatially closed space–times should fail to be singular only under exceptional circumstances."

In a more differential geometric vein, S. T. Yau in the early 1980's (unpublished) proposed the idea of a "rigid singularity theorem," to use the language of Galloway (1993). This was later stated in Bartnik (1988b) as follows:

Conjecture. *Let (M, g) be a space–time of dimension greater than two which*

(1) *contains a compact Cauchy surface, and*
(2) *satisfies the timelike convergence condition* $\mathrm{Ric}(v, v) \geq 0$ *for all timelike vectors* v.

Then either (M, g) is timelike geodesically incomplete, or else (M, g) splits isometrically as a product $(\mathbb{R} \times V, -dt^2 \oplus h)$, where (V, h) is a compact Riemannian manifold.

This philosophy apparently provided Yau with the motivation to pose in the well-known problem section of the Annals of Mathematics Studies, Volume 102 [cf. Yau (1982)], the problem of obtaining the Lorentzian analogue of the Cheeger–Gromoll Ricci Curvature Splitting Theorem stated above. Specifically, Yau posed the question as follows: Show that a space–time (M, g) which is timelike geodesically complete, obeys the timelike convergence condition, and contains a complete timelike line, splits as an isometric product $(\mathbb{R} \times V, -dt^2 \oplus h)$. This problem was then studied by Galloway (1984c) as well as by Beem, Ehrlich, and Markvorsen. Galloway considered spatially closed space–times and employed maximal surface techniques stemming from results in Gerhardt (1983) and Bartnik (1984). The approach of Beem, Ehrlich and Markvorsen, which was directed toward globally hyperbolic rather than timelike geodesically complete space–times, employed entirely different methods in which the use of the Busemann function for a future complete timelike geodesic ray was introduced into space–time geometry. Under the less stringent timelike sectional curvature assumption $K \leq 0$, a splitting theorem was obtained; combining forces with Galloway yielded Beem, Ehrlich, Markvorsen, and Galloway (1984, 1985).

A Riemannian proof contained in Eschenburg and Heintze (1984) for the Cheeger–Gromoll Riemannian Splitting Theorem, different than that originally given in Cheeger and Gromoll (1971), provided a helpful model for obtaining results in the space–time context. The d'Alembertian operator for a space–time is hyperbolic rather than elliptic, but Cheeger and Gromoll (1971) used analysis methods relying on ellipticity. Next Eschenburg (1988) made a key new observation that if one only studied the geometry in a neighborhood of the given timelike line, then the Ricci curvature assumption would allow certain arguments to be successfully modified from the $K \leq 0$ case to the Ricci curvature case, provided that the space–time was globally hyperbolic and timelike geodesically complete; then a "continuation type argument" would allow the splitting obtained in a tubular neighborhood of the given complete timelike geodesic to be extended to the entire space–time. Galloway (1989a) removed the assumption of timelike completeness from Eschenburg's work, and finally Newman (1990), rounding out the circle of papers, gave a proof assuming timelike completeness rather than global hyperbolicity. Here the new philosophy was that even though nonspacelike geodesic connectibility is not ensured without the assumption of global hyperbolicity, within a tubular neighborhood of the given complete timelike geodesic, the existence of the single complete geodesic enables certain limiting arguments to be made successfully in this tubular neighborhood despite the possible general lack of geodesic connectibility in other regions of the space–time. It should also be recalled that certain elementary consequences of the existence of a maximal timelike geodesic segment have already been treated in Section 4.4.

All of these results may be summarized in the following simple statement of the Lorentzian Splitting Theorem.

Lorentzian Splitting Theorem. *Let (M, g) be a space–time of dimension $n > 2$ which satisfies the following conditions:*

(1) *(M, g) is either globally hyperbolic or timelike geodesically complete;*
(2) *(M, g) satisfies the timelike convergence condition; and*
(3) *(M, g) contains a complete timelike line.*

Then (M, g) splits isometrically as a product $(\mathbb{R} \times V, -dt^2 \oplus h)$, where (V, h) is a complete Riemannian manifold.

Many of the arguments given in the references cited above are rather complicated. Thus in Galloway and Horta (1995), considerable simplifications were obtained by the use of "almost maximizers" to construct co-rays rather than relying on maximal geodesics as in the previous papers and in the classical constructions done for complete Riemannian manifolds [cf. Newman (1990, Lemma 3.9)]. Since this method of proof fits rather well with the spirit of Proposition 8.2 of Section 8.1, it is this approach to the Lorentzian Splitting Theorem which will be treated here.

An elementary complication in the proof of the space–time splitting theorem occurs in the construction of what would be called in Riemannian geometry an "asymptotic geodesic" to a given complete timelike geodesic through a limiting process. In the space–time case, some care is needed to ensure that the limiting geodesic, which on a priori abstract grounds could be timelike or null, turns out to be timelike rather than null.

14.1 The Busemann Function of a Timelike Geodesic Ray

We begin with a few remarks about the Riemannian Busemann function. Thus let (N, h) be an arbitrary Riemannian manifold with associated distance function $d_0 : N \times N \to \mathbb{R}$. Recall that even for incomplete Riemannian manifolds, the Riemannian distance function is *always* continuous and finite-valued. Let $c : [0, +\infty) \to (N, h)$ be any future complete unit speed geodesic *ray*, i.e., $d_0(c(0), c(t)) = t$ for all $t \geq 0$. Then we may consider a Busemann function b associated to the geodesic ray c by the formula

(14.6)
$$\begin{aligned} b(q) &= \lim_{t\to+\infty}[t - d_0(c(t), q)] \\ &= \lim_{t\to+\infty}[d_0(c(0), c(t)) - d_0(c(t), q)]. \end{aligned}$$

Since the Riemannian distance function is continuous and the triangle inequality holds, elementary arguments reveal that $b(q)$ exists, is finite-valued, and is continuous. Let us review these arguments to have them firmly fixed. Put $f(q,t) = t - d_0(c(t), q)$, so that $b(q)$ is the limit of $f(q,t)$ as t approaches $+\infty$. First the triangle inequality yields

(14.7)
$$f(q,t) \leq d_0(c(0), q)$$

and second for any $t > s$ that

(14.8)
$$f(q,t) \geq f(q,s).$$

Thus for q fixed, the function $t \mapsto f(t,q)$ is an increasing function which, by (14.7) and the finiteness of Riemannian distance, is bounded from above hence has a limit. Further, one has the elementary estimate

$$|f(p,t) - f(q,t)| \leq d_0(p,q) \tag{14.9}$$

which shows the equicontinuity of this family in t and q and passes in the limit to the inequality

$$|b(p) - b(q)| \leq d_0(p,q) \tag{14.10}$$

which evidently establishes the continuity of the Riemannian Busemann function.

Recall that for non-globally hyperbolic space–times, the Lorentzian distance function need not be continuous, only lower-semicontinuous (cf. Lemma 4.4), and also that chronologically related pairs of points $p \ll q$ may have $d(p,q) = +\infty$ (cf. Figure 4.2 and Lemma 4.2). Further, $d(p,q) = 0$ automatically if q is not contained in $J^+(p)$, and the triangle inequality is replaced by the reverse triangle inequality, assuming the points involved satisfy appropriate causality relations. Thus some care is needed in the preliminary analysis of the space–time Busemann function, especially as one case of the Lorentzian splitting problem assumes timelike geodesic completeness and *not* global hyperbolicity. Even assuming global hyperbolicity and hence guaranteeing a continuous, finite-valued distance function, examples of space–times conformal to Minkowski space show that it is possible for the space–time Busemann function to assume the value $-\infty$ or to be discontinuous unless further assumptions are made. In Beem, Ehrlich, Markvorsen, and Galloway (1985), which studied the Lorentzian splitting problem for globally hyperbolic space–times with timelike sectional curvatures $K \leq 0$, it was shown that both Busemann functions b^+ and b^- associated with the two different ends of a complete timelike geodesic line c were continuous on the subset $I(c) = I^-(c) \cap I^+(c)$, and it was also established that $I(c) = M$. Hence the sectional curvature assumption ensured that the Busemann functions associated with a complete timelike line were continuous on all of M.

In subsequent work as summarized in the introduction, in which various versions of the splitting theorem with the more desirable timelike convergence Ricci curvature condition were established, an important new realization was first found in Eschenburg (1988). This was the realization that with the weaker curvature condition, although regularity for the Busemann functions could not initially be obtained on large subsets of the given manifold, the existence of the future complete timelike line ensured good behavior of the Busemann function and of asymptotic geodesics in a tubular neighborhood of the given line. This control also sufficed to prove the Lorentzian Splitting Theorem for globally hyperbolic, timelike geodesically complete space–times. For instance, in Section 3 of Eschenburg (1988) it was shown that the Busemann function is Lipschitz continuous in a neighborhood of a given timelike geodesic ray.

An important aspect of the differential geometry of a complete, noncompact Riemannian manifold (N, g_0) is the construction of an asymptotic geodesic ray, starting at any p in N, to a given future complete geodesic ray $c : [0, +\infty) \to (N, g_0)$. Let us review this construction to motivate some of the technicalities which must be overcome in the space–time setting. Take $\{t_n\}$ with $t_n \to +\infty$. By the Hopf–Rinow Theorem, there exists a minimal unit speed geodesic c_n with $c_n(0) = p$ and $c_n(d_0(p, c(t_n))) = c(t_n)$. Letting v be any accumulation point of the sequence of unit tangent vectors $\{c_n{}'(0)\}$ in T_pM, put $\gamma(t) = \exp_p(tv)$, defined for all $t \geq 0$ by the geodesic completeness. As a limit of a subsequence of the minimal geodesic segments $\{c_n \mid [0, d_0(p, c(t_n))]\}$, c must be globally minimal, recalling that

$$\lim d_0(p, c(t_n)) = +\infty \tag{14.11}$$

since the triangle inequality gives the immediate estimate

$$d_0(p, c(t_n)) \geq d_0(c(0), c(t_n)) - d_0(p, c(0)) = t_n - d_0(p, c(0)).$$

Even though S. T. Yau formulated the Lorentzian splitting problem for timelike geodesically complete space–times, as recalled in the introduction to this chapter, the first attacks on this problem were carried out instead for globally hyperbolic space–times, for this is the class of space–times for which maximal geodesic segments exist connecting causally related pairs of points. Hence, one may at least start the construction as above by taking maximal timelike unit speed segments $c_n : [0, d(p, c(t_n))] \to (M, g)$.

Now, however, a problem arises in considering $\{c_n'(0)\}$ in that even though the space of future nonspacelike directions at p is compact (despite the fact that the set of unit future timelike tangent vectors is not compact), the limit direction v obtained for this sequence might be a *null* vector rather than a timelike vector. Dealing with this technicality motivated the introduction of the timelike co-ray condition in Beem, Ehrlich, Markvorsen, and Galloway (1985) as well as the proof that this condition would be satisfied provided that the timelike sectional curvature assumption $K \leq 0$ was imposed in order to bring Harris's Toponogov Triangle Theorem of Appendix A to bear.

Subsequent proofs of the Lorentzian Splitting Theorem in the globally hyperbolic case under the weaker Ricci curvature assumption dealt with showing that the asymptotic geodesic construction would be well behaved in some neighborhood of the given timelike ray. On the other hand, when Newman (1990) returned to Yau's original formulation of the problem, he had to deal with the further complication that the existence of maximal geodesic segments connecting pairs of causally related points could not be assumed. This led to the necessity of working with sequences of "almost maximal curves" as in Section 8.1 and also in the treatment of Galloway and Horta (1995) to the consideration of the "generalized co-ray condition."

In the construction of asymptotic geodesics from almost maximal curves as treated in Galloway and Horta (1995), a somewhat different formulation of the basic nonspacelike limit curve apparatus is employed than that discussed in Section 3.3 above, particularly Propositions 3.31 and 3.34. Their approach has the advantage of not requiring the assumption of strong causality since the topology of uniform convergence on compact subsets, rather than the C^0 topology on curves, is used in Proposition 14.3 below.

Fix a complete Riemannian metric h for the space–time (M, g) throughout the rest of this section.

Convention 14.1. Given a future inextendible nonspacelike curve γ in (M, g), γ will always be assumed to be *reparametrized* to be an h-unit speed curve unless otherwise stated.

Hence, with such a parametrization, the Hopf–Rinow Theorem guarantees that

$$\gamma : [a_0, +\infty) \to (M, g)$$

(cf. Lemma 3.65). Similarly, a future causal curve c which is both past and future inextendible may always be given an h-parametrization

$$c : (-\infty, +\infty) \to (M, g).$$

Starting with Galloway (1986a) and continuing in Eschenburg and Galloway (1992) and Galloway and Horta (1995), the following formulation of the Limit Curve Lemma has been found helpful.

Lemma 14.2 (Limit Curve Lemma). *Let $\gamma_n : (-\infty, +\infty) \to (M, g)$ be a sequence of causal curves parametrized with respect to h-arc length, and suppose that $p \in M$ is an accumulation point of the sequence $\{\gamma_n(0)\}$. Then there exist an inextendible future causal curve $\gamma : (-\infty, +\infty) \to (M, g)$ such that $\gamma(0) = p$ and a subsequence $\{\gamma_m\}$ which converges to γ uniformly with respect to h on compact subsets of $\mathbb{R}$.*

As noted in Eschenburg and Galloway (1992), an advantage to the use of the h-limit curve reparametrization is the following upper semicontinuity result, obtained without the assumption of strong causality (cf. Remark 3.35).

Proposition 14.3. *If a sequence $\gamma_n : [a, b] \to (M, g)$ of future causal curves (parametrized with h-arc length) converges uniformly to the causal curve $\gamma : [a, b] \to (M, g)$, then*

$$L(\gamma) \geq \limsup L(\gamma_n). \tag{14.12}$$

Proof. Partition $[a, b]$ as $a = t_0 < t_1 < \cdots < t_n = b$ such that each subsegment $\gamma \,|\, [t_i, t_{i+1}]$ is contained in a convex normal neighborhood N_i. By the assumption of uniform convergence, $\gamma_n \mid [t_i, t_{i+1}]$ is also contained in this neighborhood N_i for all sufficiently large n. Hence the known upper-semicontinuity of the Lorentzian arc length functional may be applied to the strongly causal space–time $(N_i, g\,|_{N_i})$ to obtain $L(\gamma \mid [t_i, t_{i+1}]) \geq \limsup_{n\to+\infty} L(\gamma_n \mid [t_i, t_{i+1}])$, and now summing over i produces inequality (14.12). □

As was noted in Eschenburg and Galloway (1992, pp. 211–212), a second advantage of the use of the auxiliary h-parameter is that the Busemann function may be defined for a future incomplete timelike geodesic ray $\gamma : [0, a) \to (M, g)$ even though the finiteness of a does not permit $t \to +\infty$ in the expression analogous to (14.6) above. But if we reparametrize γ with h-arc length, then $\gamma : [0, +\infty) \to (M, g)$, and we may put

$$b(q) = \lim_{r\to\infty} b_r(q) \tag{14.13}$$

where

$$b_r(q) = d(\gamma(0), \gamma(r)) - d(q, \gamma(r)) \tag{14.14}$$

and where d denotes the *Lorentzian* distance function of the given space–time (M, g). Note that if $q \in M - I^-(\gamma)$, then $d(q, \gamma(r)) = 0$. Hence if γ is future complete and thus $a = +\infty$, we have that $b(q) = +\infty$. Thus the space–time Busemann function $b : M \to [-\infty, +\infty]$ in general.

As recalled above, for general nonglobally hyperbolic space–times, the Lorentzian distance function can exhibit various pathologies. Nonetheless, as first noted in Eschenburg (1988), the existence of a timelike geodesic ray γ gives finiteness of the distance function and Busemann function in $I^-(\gamma) \cap I^+(\gamma(0))$. It will be useful to develop these properties in the more general context of S-rays, as introduced in Eschenburg and Galloway (1992). Thus let S be a subset of M, and recall that the distance from q to S is defined as $d(S, q) = \sup\{d(p, q) : p \in S\}$.

Definition 14.4. (*S-ray in a Space–time*) The future inextendible causal curve γ with h-arc length parametrization $\gamma : [0, +\infty) \to (M, g)$ is said to be an S-ray if $\gamma(0) \in S$ and γ maximizes distance to S, i.e.,

$$L(\gamma \mid [0, a]) = d(S, \gamma(a)) \quad \text{for all } a \geq 0. \tag{14.15}$$

A future directed nonspacelike geodesic ray in the sense of Definition 8.8 may then be given an h-reparametrization as an S-ray with S taken to be $S = \{\gamma(0)\}$. Conversely, an S-ray γ in the sense of Definition 14.4 may be reparametrized to be a nonspacelike geodesic ray in the sense of Definition 8.8 because

$$d(\gamma(0), \gamma(t)) \leq d(S, \gamma(t)) = L(\gamma \,|\, [0, t]),$$

which implies that γ may be reparametrized as a Lorentzian distance realizing geodesic and $d(\gamma(0), \gamma(t))$ is finite for any $t \geq 0$. Of course, in the applications S will usually be taken to be a spacelike hypersurface unless $S = \{\gamma(0)\}$.

The first steps toward establishing the regularity of the Busemann function and the Lorentzian distance function in a neighborhood of the given timelike ray may now be taken [cf. Eschenburg and Galloway (1992), Galloway and Horta (1995)].

Lemma 14.5. *Let $\gamma : [0, +\infty) \to (M, g)$ be an S-ray parametrized by h-arc length. Then*

(1) *$d(p, q) < +\infty$ for all $p, q \in I^-(\gamma) \cap I^+(S)$, and especially $d(p, p) = 0$ for all $p \in I^-(\gamma) \cap I^+(S)$;*

(2) *For all $q \in I^-(\gamma) \cap I^+(S)$, $d(S, q) > 0$ is finite;*

(3) *The Busemann function $b : I^-(\gamma) \to [-\infty, +\infty]$ associated to γ exists and is upper semicontinuous;*

(4) *The Busemann function b associated to the S-ray γ is finite-valued and positive on $I^-(\gamma) \cap I^+(S)$;*

(5) *For $p, q \in I^-(\gamma) \cap I^+(S)$ with $p \leqslant q$, the Busemann function satisfies $b(q) \geq b(p) + d(p, q)$; and*

(6) *Suppose $q_n = \gamma(t_n)$ and $p_n = \gamma(s_n)$ with $s_n \leq t_n$, and suppose $t_n \to t$ and $s_n \to s$. Put $q = \gamma(t)$ and $p = \gamma(s)$. Then $d(p, q) = \lim d(p_n, q_n)$.*

Proof. (1) Unless $p \leqslant q$, $d(p, q) = 0$ automatically. Thus suppose $p \leqslant q$. Take any $s \in S \cap J^-(p)$, and choose $r_0 > 0$ so that $s \leqslant p \leqslant q \ll \gamma(r)$ for any such s and $r \geq r_0$. The reverse triangle inequality then yields

$$\begin{aligned} d(s, p) + d(p, q) + d(q, \gamma(r)) &\leq d(s, \gamma(r)) \\ &\leq d(S, \gamma(r)) = L(\gamma \,|\, [0, r]) < +\infty. \end{aligned}$$

In particular, $d(p,q)$ must be finite. From Lemma 4.2–(1) we then have $d(p,p) = 0$ for all $p \in I^-(\gamma) \cap I^+(S)$.

Recalling that $L(\gamma \mid [0,r]) = d(\gamma(0), \gamma(r))$, the above chain of inequalities also yields $d(s,p) + d(p,q) + d(q,\gamma(r)) \leq d(\gamma(0),\gamma(r))$ for any s in $S \cap J^-(p)$. Since $d(s,p) = 0$ automatically if $s \in S - J^-(p)$, we obtain from this last inequality

$$d(S,p) + d(p,q) + d(q,\gamma(r)) \leq d(\gamma(0),\gamma(r)) \tag{14.16}$$

for $p \leqslant q$ and any $r \geq r_0$ chosen as above.

(2) Take $p = q$ in (14.16) to obtain

$$d(S,q) + d(q,\gamma(r)) \leq d(\gamma(0),\gamma(r)) \tag{14.17}$$

for all sufficiently large r. Since $d(\gamma(0),\gamma(r))$ is finite, $d(S,q)$ must also be finite. Since $q \in I^+(S)$, $d(s,q) > 0$ for some s in S; hence $d(S,q) > 0$.

(3) Suppose $x \in I^-(\gamma)$. Thus there exists $r > 0$ with $x \ll \gamma(r)$. Take any $s > r$. Then

$$d(x,\gamma(s)) \geq d(x,\gamma(r)) + d(\gamma(r),\gamma(s)),$$

and also

$$d(\gamma(0),\gamma(s)) = d(\gamma(0),\gamma(r)) + d(\gamma(r),\gamma(s)).$$

Hence

$$\begin{aligned} b_s(x) &= d(\gamma(0),\gamma(s)) - d(x,\gamma(s)) \\ &\leq d(\gamma(0),\gamma(r)) + d(\gamma(r),\gamma(s)) - [d(x,\gamma(r)) + d(\gamma(r),\gamma(s))] \\ &= d(\gamma(0),\gamma(r)) - d(x,\gamma(r)) \\ &= b_r(x). \end{aligned}$$

Thus for fixed $x \in I^-(\gamma)$, $r \mapsto b_r(x)$ is decreasing monotonically in r, so that $b(x) = \lim_{r\to+\infty} b_r(x)$ exists, possibly with the value $-\infty$ or $+\infty$. By the lower semicontinuity of Lorentzian distance (cf. Lemma 4.4), each $b_r(x)$ is upper semicontinuous, and hence the limit function $b(x)$ is also upper semicontinuous.

(4) Given any $x \in I^-(\gamma) \cap I^+(S)$, choose $s > 0$ so that $x \ll \gamma(s)$. Then

$b_s(x) = d(\gamma(0), \gamma(s)) - d(x, \gamma(s))$ is finite by (1) applied to $p = x$, $q = \gamma(s)$, since $d(\gamma(0), \gamma(s)) = L(\gamma \,|\, [0, s])$ is automatically finite. Hence as $r \mapsto b_r(x)$ is monotone decreasing, we have $b(x) < +\infty$. It remains to rule out $b(x) = -\infty$. To this end, take $q = x$ in equation (14.17), and subtract $d(x, \gamma(r))$ to obtain $b_r(x) \geq d(S, x) > 0$ for any $r > s$. Taking limits yields

$$b(x) \geq d(S, x) > 0 \tag{14.18}$$

for any x in $I^-(\gamma) \cap I^+(S)$.
(5) Choose r_0 so that $p \leqslant q \ll \gamma(r)$ for all $r \geq r_0$. Applying the reverse triangle inequality to $p \leqslant q \ll \gamma(r)$ yields

$$\begin{aligned} b_r(p) &= d(\gamma(0), \gamma(r)) - d(p, \gamma(r)) \\ &\leq d(\gamma(0), \gamma(r)) - d(p, q) - d(q, \gamma(r)) \\ &= b_r(q) - d(p, q). \end{aligned}$$

Let $r \to +\infty$ to obtain the desired inequality.
(6) Since γ is maximal, we have $d(p_n, q_n) = L(\gamma \,|\, [s_n, t_n])$ for all n. Thus $\lim d(p_n, q_n) = \lim L(\gamma \,|\, [s_n, t_n]) = \lim L(\gamma \,|\, [s, t\,]) = d(p, q)$. □

It has been noted above that in considering the construction of asymptotic geodesics, somewhat different techniques must be employed to deal with the timelike geodesically complete, but not necessarily globally hyperbolic, case. Hence we find certain conventions employed in Galloway and Horta (1995) to be helpful and will use them below. First, we formalize certain aspects of Section 8.1, recalling here that unlike Section 8.1, strong causality is not assumed and also that curves are always given an h-arc length parametrization in this section unless otherwise stated.

Definition 14.6. (*Limit Maximizing Sequence of Causal Curves*) A sequence $\gamma_n : [a_n, b_n] \to (M, g)$ of h-arc length parametrized future causal curves is said to be *limit maximizing* if

$$L(\gamma_n) \geq d(\gamma_n(a_n), \gamma_n(b_n)) - \epsilon_n \tag{14.19}$$

where $\epsilon_n \to 0$ as $n \to +\infty$.

Given any p, q in (M, g) with $0 < d(p,q) < +\infty$, the definition of Lorentzian distance implies that a sequence of limit maximal curves from p to q may be constructed. In a manner similar to the proofs in Section 8.1, the lower semicontinuity of distance and the upper semicontinuity of L established in Proposition 14.3 combine to yield

Proposition 14.7. *Suppose that $\gamma_n : [a_n, b_n] \to (M, g)$ is a limit maximizing sequence of future directed causal curves that converges uniformly to a causal curve $\gamma : [a, b] \to M$ on some subinterval $[a, b] \subseteq \bigcap_n [a_n, b_n]$. Then $L(\gamma) = d(\gamma(a), \gamma(b))$, and hence γ may be reparametrized as a future directed maximal geodesic from $\gamma(a)$ to $\gamma(b)$.*

Proof. Putting $\gamma_n = \gamma_n \,|\, [a, b]$, we have as usual

$$\begin{aligned}\limsup_{n\to+\infty} L(\gamma_n) \le L(\gamma) &\le d(\gamma(a), \gamma(b)) \\ &\le \liminf_{n\to+\infty} d(\gamma_n(a), \gamma_n(b)) \\ &\le \liminf_{n\to+\infty} (L(\gamma_n) + \epsilon_n) \\ &= \liminf_{n\to+\infty} L(\gamma_n),\end{aligned}$$

whence

$$L(\gamma) = \lim_{n\to+\infty} L(\gamma_n) = \lim_{n\to+\infty} d(\gamma_n(a), \gamma_n(b)) = d(\gamma(a), \gamma(b)). \quad \square \tag{14.20}$$

The following lemma from Eschenburg and Galloway (1992) has been helpful in constructing asymptotic geodesic rays. It should be emphasized that unlike (14.11) above for complete Riemannian manifolds, the condition $d(z_n, p_n) \to +\infty$ in Lemma 14.8 does *not* imply that $\{p_n\}$ diverges to infinity but only that $a_n \to +\infty$.

Lemma 14.8. *Let $\{z_n\}$ be a sequence in (M, g) with $z_n \to z$. Let $p_n \in I^+(z_n)$ with $d(z_n, p_n) < +\infty$. Let $\gamma_n : [0, a_n] \to (M, g)$ be a limit maximizing sequence of future causal curves with $\gamma_n(0) = z_n$ and $\gamma_n(a_n) = p_n$. Let $\widetilde{\gamma}_n : [0, +\infty) \to (M, g)$ be any future inextendible extension of γ_n. Suppose that $d(p_n, z_n) \to +\infty$. Then any limit curve $\gamma : [0, +\infty) \to (M, g)$ of the sequence $\{\widetilde{\gamma}_n\}$ is a nonspacelike geodesic ray starting at z.*

Proof. It is necessary to show that $a_n \to +\infty$. Suppose not. By passing to a subsequence, we may suppose that $a_n \to a < +\infty$. Because the curves γ_n are parametrized by h-arc length and h is a complete Riemannian metric, it follows that all γ_n are contained in a compact subset K of M. Note also that in this case we have $L_h(\gamma_n) = a_n \to a < +\infty$.

On the other hand, we will show that the assumption that $d(z_n, p_n) \to +\infty$ implies that $L_h(\gamma_n) \to +\infty$, yielding the required contradiction. To this end, put a second auxiliary Riemannian metric h_0 on M more closely associated to the given Lorentzian metric g by the following standard construction. Let T be a unit timelike vector field on (M, g), and define an associated one-form by $\tau = g(T, \cdot)$. Put

$$h_0 = g + 2\tau \otimes \tau = g^{\perp} + \tau \otimes \tau,$$

and note that for this second Riemannian metric we have the given Lorentzian arc length of any causal curve segment c and the h_0-arc length related by

$$L_{h_0}(c) \geq L_g(c).$$

Since h and h_0 are Riemannian metrics and K is compact, there exists a constant $\lambda > 0$ such that $h(v, v) \geq \lambda^2 h_0(v, v)$ for any $v \in TM \,|\, K$, whence

$$L_h(c) \geq \lambda L_{h_0}(c)$$

for any curve segment c contained in K. Thus we have

$$L_h(\gamma_n) \geq \lambda L_g(\gamma_n) \geq \lambda d(z_n, p_n) - \lambda \epsilon_n \to +\infty \quad \text{by hypothesis.} \quad \square$$

A more complex lemma is considered in Eschenburg and Galloway (1992) in order to allow for dealing with inextendible timelike rays which are not necessarily future complete. But in view of the hypotheses involved in the Lorentzian splitting problem, let us suppose now that $\gamma : [0, +\infty) \to (M, g)$ is a *future complete* timelike S-ray, hence of infinite Lorentzian length. Fix $z \in I^-(\gamma) \cap I^+(S)$. Let $z_n \to z$, and put $p_n = \gamma(r_n)$ where $r_n \to +\infty$. Then $d(z_n, p_n) < +\infty$ for all n sufficiently large by Lemma 14.5–(1). On the other hand, if we fix $s > 0$ with $z \ll \gamma(s)$, then for all sufficiently large n we have

$z_n \ll \gamma(s)$ as well. Hence the reverse triangle inequality yields $d(z_n, p_n) \geq d(z_n, \gamma(s)) + d(\gamma(s), \gamma(r_n)) \to +\infty$ since $d(\gamma(s), \gamma(r_n)) = L(\gamma \,|\, [s, r_n]) \to +\infty$ by hypothesis.

Thus with Lemma 14.8 in hand and the above verification, the construction of asymptotic geodesics to a future complete S-ray, $\gamma : [0, +\infty) \to (M, g)$, may now be accomplished. Simply let $\{\mu_n\}$ be any sequence of limit maximizing curves from z_n to p_n, and let $\mu : [0, +\infty) \to (M, g)$ be any limit curve guaranteed by Lemma 14.8 for the extended curves $\{\widetilde{\mu}_n\}$. The use of the auxiliary complete Riemannian metric h to parametrize the curves by h-arc length is what ensures the uniform convergence on compact subintervals necessary to apply Proposition 14.7 and Lemma 14.8.

We will adopt the following terminology of Galloway and Horta (1995).

Definition 14.9. (*Co-rays, Generalized Co-rays, and Asymptotes*)

(1) Any ray μ constructed in this fashion will be called a *generalized co-ray* to the given future complete S-ray γ.

(2) The ray μ constructed in this fashion will be called a *co-ray* to the given future complete S-ray γ if all μ_n are distance maximizers, i.e., $L(\mu_n) = d(z_n, p_n)$ for all n.

(3) The ray μ constructed in this fashion will be called an *asymptote* if all μ_n are distance maximizers and if $z_n = z$ for all n.

In the space–time setting, the concept and term "co-ray" were first introduced in Beem, Ehrlich, Markvorsen, and Galloway (1985), influenced by terminology in Busemann (1955). Generalized co-rays for space–times were first employed in Eschenburg and Galloway (1992).

While considering limit maximizing sequences of timelike curves, Newman (1990, Lemma 3.9) noticed an important implication of timelike geodesic completeness which is formulated in Galloway and Horta (1995) as follows:

Proposition 14.10. *Let (M, g) be future timelike geodesically complete. Suppose p, q are points in (M, g) with $p \ll q$ and $d(p, q) < +\infty$. Let $\gamma_n : [0, a_n] \to (M, g)$ be a limit maximizing sequence of future directed causal curves from p to q. For each n, let $\widetilde{\gamma}_n : [0, +\infty) \to (M, g)$ be a future inex-*

tendible extension of γ_n. *Then if* $\gamma : [0, +\infty) \to (M, g)$ *is a limit curve of* $\{\widetilde{\gamma}_n\}$, *either* γ *maximizes from* p *to* q *or* γ *is a null ray.*

Proof. Let $a = \sup\{a_n\}$. The existence of convex neighborhoods centered at p implies $a > 0$. By passing to a subsequence if necessary, we may assume that $a_n \to a$. Then by Proposition 14.7, the limit curve $\gamma \,|\, [0, a)$ is a maximal nonspacelike geodesic.

First suppose $a < +\infty$. Because we have chosen an auxiliary complete Riemannian metric and parametrized the γ_n by h-arc length, we know that $\gamma \,|\, [0, a)$ extends to $t = a$. Also, we have that $\gamma \,|\, [0, a]$ is maximal, and $\gamma(a) = \lim_{n \to +\infty} \gamma_n(a_n) = q$. Since $d(p, q) > 0$ and γ is maximal from p to q in this case, γ must be timelike.

Now suppose $a = +\infty$. If γ is null, then γ is the required null ray, and we are finished. Thus suppose γ is timelike. By the assumption of geodesic completeness, γ has infinite Lorentzian length. As $d(p, q)$ is a finite positive number, there exists $T > 0$ such that $L(\gamma \,|\, [0, T]) > d(p, q)$. Since $a_n \to +\infty$, equation (14.20) in the proof of Proposition 14.7 implies that for n sufficiently large, we have $T < a_n$, and hence

$$L(\gamma_n \,|\, [0, a_n]) \geq L(\gamma_n \,|\, [0, T]) > d(p, q).$$

But this is impossible because γ_n is a timelike curve segment from p to q, whence $L(\gamma_n) \leq d(p, q)$ by definition of Lorentzian distance. □

14.2 Co-rays and the Busemann Function

It was first noted in Beem, Ehrlich, Markvorsen, and Galloway (1985) that the assumption for a globally hyperbolic space–time that all co-rays constructed to either end of a complete timelike geodesic line $\gamma : (-\infty, +\infty) \to (M, g)$ turn out to be timelike rather than null implies that the Busemann functions b^+ and b^- associated to the two different ends of γ are continuous and also finite-valued on the sets $I^-(\gamma)$ and $I^+(\gamma)$, respectively. This assumption was called the "timelike co-ray condition." It was also shown, employing Harris's Lorentzian version of the Toponogov Comparison Theorem for timelike geodesic triangles in globally hyperbolic space–times (cf. Appendix A), that the assumption of everywhere nonpositive timelike sectional curvatures ensures that the given globally hyperbolic space–time satisfies this timelike co-ray condition.

In the subsequent works, beginning with Eschenburg (1988), more difficult studies of the regularity properties of the Busemann function were made assuming only the timelike convergence condition on the Ricci curvature; hence the space–time Toponogov Comparison Theorem could not be brought to bear on this issue.

Another ingredient in the proof of the Lorentzian Splitting Theorems was the fortuitous publication of Eschenburg and Heintze (1984) in which a new proof method for the Riemannian Cheeger–Gromoll Ricci Curvature Splitting Theorem was employed, utilizing the technique of smooth upper and lower support functions for the (a priori only continuous) Busemann functions of a geodesic line. The original proof method of Cheeger and Gromoll (1971) relied heavily on the ellipticity of the Laplacian and hence did not seem to be adaptable to the case of the hyperbolic d'Alembertian.

We begin this section with the definition of these support functions as first treated in Eschenburg (1988) for the globally hyperbolic case and later treated in Galloway and Horta (1995) in the current context of timelike geodesic completeness. Let $\gamma : [0, +\infty) \to (M, g)$ be a *future complete* S-ray in a space–time (M, g), and let b be the associated Busemann function given by equations (14.13) and (14.14).

Definition 14.11. (*Upper Support Function*) Let $\alpha : [0, +\infty) \to (M, g)$ be a timelike asymptote to γ with $\alpha(0) = p \in I^-(\gamma) \cap I^+(S)$. For each $s > 0$, let

$$b_{p,s} : M \to [-\infty, +\infty]$$

be defined by

$$b_{p,s}(x) = b(p) + d(p, \alpha(s)) - d(x, \alpha(s)). \tag{14.21}$$

Lemma 14.12. *For any $s > 0$ the function $b_{p,s}$ given by (14.21) is a continuous upper support function for b at p, i.e., $b_{p,s}(x) \geq b(x)$ for all x near p with equality holding when $x = p$.*

Proof. By definition of an asymptote, there exists a sequence of maximal timelike geodesic segments $\alpha_n : [0, a_n] \to (M, g)$ with $\alpha_n(0) = p$ and $\alpha_n(a_n) = \gamma(r_n)$, where $r_n \to +\infty$ as $n \to +\infty$. Also, since γ is future complete, $a_n \to +\infty$ since $d(p, \gamma(r_n)) \to +\infty$. Let x be in the open neighborhood $U = I^-(\alpha(s)) \cap I^+(S)$ of p for the fixed $s > 0$. Then for n sufficiently large, $x \ll \alpha_n(s)$ as well. By lower semicontinuity of distance, there exists $\delta_n \to 0$ such that $d(x, \alpha_n(s)) \geq d(x, \alpha(s)) - \delta_n$ for all large n. Then for all n sufficiently large, the reverse triangle inequality and the maximality of γ imply

(14.22)
$$\begin{aligned} b_{r_n}(x) - b_{r_n}(p) &= d(p, \gamma(r_n)) - d(x, \gamma(r_n)) \\ &\leq d(p, \gamma(r_n)) - d(x, \alpha_n(s)) - d(\alpha_n(s), \gamma(r_n)) \\ &= d(p, \alpha_n(s)) - d(x, \alpha_n(s)) \\ &\leq d(p, \alpha_n(s)) - d(x, \alpha(s)) + \delta_n. \end{aligned}$$

By (14.20) also, $d(p, \alpha_n(s)) \to d(p, \alpha(s))$ as $n \to +\infty$. Hence taking the limit in (14.22) yields $b(x) - b(p) \leq d(p, \alpha(s)) - d(x, \alpha(s))$, so that

$$b_{p,s}(x) = b(p) + d(p, \alpha(s)) - d(x, \alpha(s)) \geq b(x) \tag{14.23}$$

as required. □

As noted by Eschenburg (1988) in the globally hyperbolic case, the expected behavior of the Busemann function associated to the given ray γ on rays α asymptotic to γ, akin to that previously observed for complete Riemannian manifolds, may now be obtained.

Corollary 14.13. *Let $\gamma : [0, +\infty) \to (M, g)$ be a future complete S-ray in the space–time (M, g), and let b be the Busemann function associated to γ. If $\alpha : [0, +\infty) \to (M, g)$ is a timelike asymptote to γ starting at p in $I^-(\gamma) \cap I^+(S)$, then $b(\alpha(t)) = d(p, \alpha(t)) + b(p)$ for all $t \geq 0$.*

Proof. Taking $x = \alpha(t)$ and fixing any $s > t$ in (14.23) gives

$$\begin{aligned} b(\alpha(t)) &\leq b(p) + d(p, \alpha(s)) - d(\alpha(t), \alpha(s)) \\ &= b(p) + d(p, \alpha(t)) \end{aligned}$$

since α is maximal. On the other hand, applying Lemma 14.5–(5) to p and $q = \alpha(t)$ yields

$$b(\alpha(t)) \geq b(p) + d(p, \alpha(t)). \quad \square$$

Now we are ready to state the crucial new condition in Galloway and Horta (1995) which leads to the local regularity of the Busemann function as well as some control over maximal geodesic connectibility.

Definition 14.14. (*Generalized Timelike Co-ray Condition at p*) Let (M, g) be a future timelike geodesically complete space–time, and let γ be a timelike S-ray. Suppose that $p \in I^-(\gamma) \cap I^+(S)$. Then the *generalized timelike co-ray condition holds at p* if every generalized co-ray to γ starting at p is timelike.

Arguments similar to those of Lemma 14.15 below show that this condition is an open condition, i.e., if the condition holds at a given $p \in I^-(\gamma) \cap I^+(S)$, then the condition holds in some neighborhood of p.

Lemma 14.15. *Suppose (M, g) is future timelike geodesically complete, and let γ be a timelike S-ray. Assume that the generalized timelike co-ray condition holds at $p \in I^-(\gamma) \cap I^+(S)$. Then there exist a neighborhood U of p and a constant $R > 0$ such that for all $q \in U$ and for all $r > R$, there exists a maximal timelike geodesic segment from q to $\gamma(r)$.*

Proof. Assuming that the desired conclusion is false, we will show that the generalized timelike co-ray condition cannot hold at the given point p by a diagonalizing argument.

Supposing that the conclusion of the lemma is false, we may find $\{p_n\} \subseteq I^-(\gamma) \cap I^+(S)$ with $p_n \to p$, $r_n \to +\infty$, and $p_n \ll \gamma(r_n)$ such that there is no maximal timelike geodesic segment from p_n to $\gamma(r_n)$ for all n. In view of Lemma 14.5–(1), we have $0 < d(p_n, \gamma(r_n)) < +\infty$ for each n. Hence for each n we may construct a limit maximizing sequence of future nonspacelike curves

$$\alpha_{nk} : [0, a_{nk}] \to (M, g)$$

from $p_n = \alpha_{nk}(0)$ to $\gamma(r_n) = \alpha_{nk}(a_{nk})$ with

$$L(\alpha_{nk}) \geq d(\alpha_{nk}(0), \alpha_{nk}(a_{nk})) - \epsilon_{nk}$$

where $\lim_{k\to+\infty} \epsilon_{nk} = 0$. Because of the assumed nonexistence of timelike maximizers from p_n to $\gamma(r_n)$, the proof of Proposition 14.10 of Section 14.1 implies that $\lim_{k\to+\infty} a_{nk} = +\infty$ and also that the sequence $\{\widetilde{\alpha}_{nk}\}$ converges to a null ray as $k \to +\infty$, which we will call $\beta_n : [0, +\infty) \to (M, g)$, with initial point $\beta_n(0) = p_n$. Hence equation (14.20) forces

$$\lim_{k\to+\infty} L(\alpha_{nk} \mid [0, s]) = L(\beta_n \mid [0, s]) = 0$$

for any $s > 0$.

Now do a diagonalizing procedure on the nonspacelike curves $\{\alpha_{nk}\}$ to find an increasing sequence $k(n) \to +\infty$ with the property that for $\eta_n = \alpha_{nk(n)}$ and $b_n = a_{nk(n)}$, the associated sequence $\{\eta_n\}$ satisfies

(1) $b_n \to +\infty$;

(2) $L(\eta_n \mid [0, 1]) < 1/n$ for all n; and

(3) $L(\eta_n \mid [0, b_n]) \geq d(\eta_n(0), \eta_n(b_n)) - 1/n$ for all n.

By (1), (3), and Proposition 14.7, the sequence $\{\eta_n\}$ converges to a nonspacelike geodesic ray η with $\eta(0) = p$. By condition (2), the ray η must be null.

Thus negating the desired conclusion, we have produced a generalized co-ray η to γ with $\eta(0) = p$ which is a null ray. But this contradicts the hypothesis of the lemma that all co-rays to γ at p are timelike. □

The next several results from Galloway and Horta (1995) treat the relationship between limit curve convergence and convergence by initial tangent direction for timelike geodesic segments and in particular the property that the timelike co-ray condition ensures that initial tangents of asymptotes constructed in a neighborhood of p are bounded away from the null cone. Out of these considerations, among others, the local continuity of the Busemann function may then be established in Theorem 14.19.

It is helpful to adopt some mechanism to produce appropriate compact subsets of the noncompact set of unit timelike tangent vectors. To be consistent with Chapter 9, we will recall the notation $T_{-1}M|_p$ from Definition 9.2 and introduce a new notation following Galloway and Horta of

$$K_C(p) = \{v \in T_{-1}M|_p : h(v, v) \leq C\} \tag{14.24}$$

where $C > 0$ and h is the auxiliary Riemannian metric for (M, g) fixed above. These sets are compact and nonempty for C sufficiently large, and evidently

$$\bigcup_{C>0} K_C(p) = T_{-1}M\big|_p .$$

Lemma 14.16. *Let (M, g) be future timelike geodesically complete, and let γ be a timelike S-ray. Assume that the generalized timelike co-ray condition holds at $p \in I^-(\gamma) \cap I^+(S)$. Then there exist a neighborhood U of p and constants $R > 0$, $C > 0$ such that for all q in U and any $r > R$, if $\alpha : [0, a] \to (M, g)$ is any maximal timelike geodesic segment from q to $\gamma(r)$ parametrized with respect to arc length in the given Lorentzian metric, then $\alpha'(0) \in K_C(q)$.*

Proof. Suppose that the lemma fails to hold. Then there exist sequences $p_n \to p$, $r_n \to +\infty$, and maximal timelike geodesics $\alpha_n : [0, a_n] \to (M, g)$ from p_n to $\gamma(r_n)$ parametrized as g-unit speed timelike curves with the further property that $h({\alpha_n}'(0), {\alpha_n}'(0)) \to +\infty$ as $n \to +\infty$. To apply the limit curve machinery, reparametrize α_n to an h-unit speed curve, denoted by $\widetilde{\alpha}_n$, with $\widetilde{\alpha}_n(0) = p_n$. Using the limit curve machinery of Section 14.1 and Lemma 14.8 produces a nonspacelike geodesic ray $\widetilde{\alpha}$ with $\widetilde{\alpha}(0) = p$ and with $\widetilde{\alpha}$ a co-ray to γ, to which a subsequence of the $\widetilde{\alpha}_n$'s converge, which we may assume is the given sequence $\{\widetilde{\alpha}_n\}$ itself. Since the generalized timelike co-ray condition is assumed to hold at p, we have that $\widetilde{\alpha}$ is timelike. Reparametrize to a g-unit speed timelike ray $\alpha : [0, +\infty) \to (M, g)$. The domain is as indicated because of the assumption of future timelike completeness.

Choose $\delta > 0$ so that $\widetilde{\alpha}([0, \delta])$ is contained in a convex normal neighborhood of p and also is contained in $I^-(\gamma)$. Because of this supposition, $\epsilon = d(\widetilde{\alpha}(0), \widetilde{\alpha}(\delta)) > 0$ and ϵ is also finite by Lemma 14.5–(1). Put $\epsilon_n = d(\widetilde{\alpha}_n(0), \widetilde{\alpha}_n(\delta))$. By equation (14.20) we have $\epsilon_n \to \epsilon$. Now using the future timelike completeness and choice of convex normal neighborhood, we may calculate

$$\epsilon_n {\alpha_n}'(0) = \exp_{p_n}^{-1}(\widetilde{\alpha}(\delta)) \to \exp_p^{-1}(\widetilde{\alpha}(\delta)) = \epsilon\, \alpha'(0).$$

Hence, ${\alpha_n}'(0) \to \alpha'(0)$ which implies that

$$h(\alpha'(0), \alpha'(0)) = \lim_{n \to +\infty} h({\alpha_n}'(0), {\alpha_n}'(0)).$$

But this is contradictory since $h(\alpha'(0), \alpha'(0))$ is finite, yet by supposition, $\lim_{n \to +\infty} h({\alpha_n}'(0), {\alpha_n}'(0)) = +\infty$. □

Corollary 14.17. *Let (M, g) be future timelike geodesically complete, and let γ be a timelike S-ray. Suppose the generalized timelike co-ray condition holds at $p \in I^-(\gamma) \cap I^+(S)$. Then there exist a neighborhood U of p and a constant $C > 0$ such that for all $q \in U$, if $\alpha : [0, +\infty) \to (M, g)$ is a co-ray (or asymptote) to γ from q, parametrized with respect to Lorentzian arc length, then $\alpha'(0) \in K_C(q)$.*

The generalized timelike co-ray condition also has implications for the continuity of the Lorentzian distance function, as noted in Galloway and Horta (1995).

Proposition 14.18. *Let (M, g) be future timelike geodesically complete, and let γ be a timelike S-ray. Suppose the generalized timelike co-ray condition holds at $p \in I^-(\gamma) \cap I^+(S)$. Then there exist a neighborhood U of p and a number $R > 0$ such that for all $r > R$ the function $\delta : U \to [0, +\infty)$ defined by $\delta(x) = d(x, \gamma(r))$ is continuous on U.*

Proof. Recall first that Lemma 14.5–(1) guarantees that δ is finite-valued on U. By Lemma 4.4, δ is lower semicontinuous, so it only remains to establish the upper semicontinuity.

Choose U, R, and C to satisfy the conclusions of Lemmas 14.15 and 14.16. Fix any $r > R$. Suppose δ is not upper semicontinuous at x in U. Then there exists a sequence $\{x_n\} \subseteq U$ with $x_n \to x$ but $\lim_{n \to +\infty} d(x_n, \gamma(r)) > d(x, \gamma(r))$. For each n, let α_n be a maximal timelike geodesic segment which is parametrized by g-arc length, from x_n to $\gamma(r)$. In view of the conditions satisfied by U, R, and C, the initial tangents $\{{\alpha_n}'(0)\}$ are contained in a compact subset of $T_{-1}M|_x$. By this fact and the assumption of timelike future completeness, there exists a subsequence $\{\alpha_{n(k)}\}$ which converges to a maximal timelike geodesic α from x to $\gamma(r)$. But then by equation (14.20), we have

$$d(x_{n(k)}, \gamma(r)) = L(\alpha_{n(k)}) \to L(\alpha) = d(x, \gamma(r)),$$

in contradiction. □

With these results now established, we are ready for the proof given in Galloway and Horta (1995) of the continuity of the Busemann function in the presence of the generalized timelike co-ray condition. The treatment is inspired by that in Eschenburg (1988) for the globally hyperbolic case.

Theorem 14.19. *Let (M, g) be future timelike geodesically complete, and let γ be a timelike S-ray. Assume that the generalized timelike co-ray condition holds at p in $I^-(\gamma) \cap I^+(S)$. Then the Busemann function b associated to γ is Lipschitz continuous on a neighborhood of p.*

Proof. It is helpful to use a lemma given in the Appendix to Eschenburg (1988).

Sublemma 14.20. *Let U be an open convex domain in $\mathbb{R}^n$ and $f : U \to \mathbb{R}$ a continuous function. Suppose that for each q in U there is a smooth lower support function f_q defined in a neighborhood of q such that $\|d(f_q)_q\| < L$. Then f is Lipschitz continuous with Lipschitz constant L, i.e., for all x, y in U we have*

$$|f(x) - f(y)| \leq L\,\|x - y\|. \tag{14.25}$$

Proof of Sublemma. First treat the case that $n = 1$ and thus U is an open interval and the estimate amounts to supposing that

$$|f_q{}'(q)| \leq L \tag{14.26}$$

for all q in U. Suppose that (14.25) is not valid. Then there exist $x < y$ in U such that by replacing f with $-f$ if necessary, we may suppose that x, y, and f satisfy $f(x) < f(y)$ and also $|f(x) - f(y)| > L\,|x - y|$. Graphical considerations evidently enable one to find an affine function $l(x)$ with constant slope $l'(x) = L_0 > L$ such that $f(x) < l(x)$ and $f(y) > l(y)$. Put $q_0 = \sup\{t \in [x, y] : f(t) < l(t)\}$. By continuity of f, evidently $f_{q_0}(q_0) = f(q_0) = l(q_0)$ and $f_{q_0}(t) \leq f(t) < l(t)$ for all $t < q_0$ sufficiently close to q_0. Hence, $f'_{q_0}(q_0) \geq l'(q_0) = L_0 > L$, in contradiction to inequality (14.26). Thus the desired estimate (14.25) must be valid in the case that $n = 1$.

Now consider the case that $n > 1$. Fixing arbitrary x, y in U, let $c(t) = (1-t)x + ty$ and $I = \mathbb{R} \cap c^{-1}(U)$. Put $g = f \circ c : I \to \mathbb{R}$. Note that in view of the chain rule, the relevant constant in (14.26) applied to g is $L_1 = L\,\|x-y\|$. Applying the $n = 1$ case to $g(t)$ with $t = 0$ and $t = 1$ yields

$$|f(x) - f(y)| = |g(0) - g(1)| \leq L_1\,|0-1| = L\,\|x-y\|$$

as required. □

Proof of Theorem 14.19. Let γ be parametrized by Lorentzian arc length below. Choose U, R, and C according to Lemmas 14.15 and 14.16. Further, take U sufficiently small that $\overline{U}$ is compact and is contained in a convex normal neighborhood (N, ϕ) such that $\phi(U)$ is convex in $\mathbb{R}^n$, $n = \dim M$. Let $\phi = (x_1, \ldots, x_n)$, and let h_0 denote the associated Euclidean metric for TN given by

$$h_0 = \sum_{i=1}^{n} dx_i{}^2.$$

Since U has compact closure and h, h_0 are both Riemannian metrics, there is a constant $K > 0$ such that $h_0 \leq K\,h$ on TU.

Put $d_r(x) = d(x, \gamma(r))$. It is sufficient to show that the functions $b_r = r - d_r$ are Lipschitz continuous on U with the same Lipschitz constant for all $r > R$. Since

$$b_r(x) - b_r(y) = d_r(y) - d_r(x),$$

this is equivalent to obtaining the Lipschitz continuity of the d_r, $r > R$, on U. From Proposition 14.18 we have the continuity of the d_r's; it thus remains only to obtain the Lipschitz estimate using Sublemma 14.20.

Let $D : N \times N \to [0, +\infty)$ denote the local Lorentzian distance function for $(N, g\,|_N)$ as described in Definition 4.25 and Lemma 4.26. Fix any point q in U and $r > R$. Let α be a maximal timelike geodesic segment from q to $\gamma(r)$ parametrized by g-arc length, guaranteed by Lemma 14.15. Fix any q' in U which lies on α with $q \neq q'$. Having made these choices, a candidate for a local smooth support function for d_r near q may then be defined by

(14.27) $$f_{q,r}(x) = D(x, q') + d(q', \gamma(r)).$$

For x in a neighborhood of q, the functions $x \mapsto D(x, q')$, and hence $f_{q,r}(x)$, are smooth by Lemma 4.26. Also the local distance function D generally satisfies $D(x, q') \leq d(x, q')$ for any x, q' in N. Thus by the reverse triangle inequality,

$$\begin{aligned} f_{q,r}(x) &\leq d(x, q') + d(q', \gamma(r)) \\ &\leq d(x, \gamma(r)) = d_r(x), \end{aligned}$$

and also

$$f_{q,r}(q) = D(q, q') + d(q', \gamma(r)) = d(q, q') + d(q', \gamma(r)) = d_r(q)$$

using the maximality of α.

It remains to establish an estimate for the lower support functions $f_{q,r}$ of the form needed to employ Sublemma 14.20. To this end, put

$$G = \sup\{|g_{ij}(x)| : x \in U, \ \ 1 \leq i, j \leq n\}$$

and $\|v\|_0 = \sqrt{h_0(v, v)}$ for $v \in T_qM$. Since U has compact closure in N, we have $G > 0$. The basic differential geometry of the distance function $d(\,\cdot\,, q')$ from the point q' prior to the past timelike cut locus of q' provides the gradient calculation

$$v\,(f_{q,r}) = g(v, \alpha'(0))$$

for any $v \in T_qM$ (cf. Lemma 14.26 above). Hence,

$$\begin{aligned} |d\,(f_{q,r})\,(v)| &= |g(v, \alpha'(0))| \\ &\leq G\,\|\alpha'(0)\|_0\,\|v\|_0 \\ &\leq GK\sqrt{h(\alpha'(0), \alpha'(0))}\,\|v\|_0 \\ &\leq GKC^{\frac{1}{2}}\,\|v\|_0. \end{aligned}$$

Hence, $L = GKC^{\frac{1}{2}}$ has the property that $\|d\,(f_{q,r})_q\| \leq L$ for all $q \in U$. Thus d_r and hence b_r are Lipschitz continuous on U with Lipschitz constant $L = GKC^{\frac{1}{2}}$ by Sublemma 14.20 for any $r > R$. $\square$

Having presented four results on the pleasing consequences of the generalized timelike co-ray condition, let us now show as in Galloway and Horta (1995) that in the important special case where $S = \{\gamma(0)\}$, and hence an S-ray is an ordinary geodesic ray, the generalized timelike co-ray condition holds in a neighborhood of the given ray, possibly excluding the initial point $\gamma(0)$. This may be accomplished by first proving the following result.

Proposition 14.21. *Let (M, g) be future timelike geodesically complete, and let $\gamma : [0, +\infty) \to (M, g)$ be a future directed timelike ray. Then any generalized co-ray starting at $p = \gamma(a)$, $a > 0$, must coincide with γ.*

Proof. Here the limit curve techniques of Section 14.1 will be employed, so all curves will be given a unit speed parametrization with respect to the fixed auxiliary Riemannian metric h. Let $\alpha : [0, +\infty) \to (M, g)$ be a generalized co-ray to γ at $p = \gamma(a)$. Thus there exists a limit maximizing sequence of future directed causal curves $\alpha_n : [0, a_n] \to (M, g)$ from p_n to $\gamma(r_n)$ with $p_n \to p$, $r_n \to +\infty$, and $a_n \to +\infty$, which converges to α. Let U be any convex normal neighborhood of p. Choose any q in U lying on γ to the past of p, and take a sequence $\{q_n\}$ of points on γ between q and p such that $q_n \to p$. Since $q_n \ll p$ and $p_k \to p$, by taking a subsequence of $\{p_n\}$ if necessary, we may suppose that $q_n \ll p_n$ for all n and that $\{p_n\} \subseteq U$. Since p_n, q_n are contained in U and $\lim p_n = \lim q_n = p$, we may connect q_n to p_n by a timelike geodesic segment whose length approaches 0 as $n \to +\infty$.

Now define a new sequence $\{\sigma_n : [0, s_n] \to (M, g)\}$ of causal curves by following γ from q to q_n, then going along the above timelike geodesic segment in U from q_n to p_n, and finally following along α_n from p_n to $\gamma(r_n)$. Put $q_n = \sigma_n(a_n)$ and $p_n = \sigma_n(b_n)$. Fix $r > a$ so that $p_n \in I^-(\gamma(r))$ and $r_n > r$ for all sufficiently large n. We now need to establish the limit maximality of this new sequence $\{\sigma_n\}$.

Since γ is maximal and the α_n's are limit maximizing, we have

$$\begin{aligned} L(\sigma_n \mid [0, s_n]) &= L(\sigma_n \mid [0, a_n]) + L(\sigma_n \mid [a_n, b_n]) + L(\sigma_n \mid [b_n, s_n]) \\ &\geq d(q, q_n) + L(\sigma_n \mid [a_n, b_n]) + d(p_n, \gamma(r_n)) - \epsilon_n \\ &= d(q, p) - d(q_n, p) + L(\sigma_n \mid [a_n, b_n]) + d(p_n, \gamma(r_n)) - \epsilon_n. \end{aligned}$$

On the other hand, the lower semicontinuity of distance at $(p, \gamma(r))$ implies the following inequality is valid:

$$\begin{aligned} d(p_n, \gamma(r_n)) &\geq d(p_n, \gamma(r)) + d(\gamma(r), \gamma(r_n)) \\ &= d(p_n, \gamma(r)) + [d(p, \gamma(r_n)) - d(p, \gamma(r))] \\ &\geq [d(p, \gamma(r)) - \delta_n] + d(p, \gamma(r_n)) - d(p, \gamma(r)) \\ &= d(p, \gamma(r_n)) - \delta_n \quad \text{with } \delta_n \to 0 \text{ as } n \to +\infty. \end{aligned}$$

Combining these two inequalities yields

$$\begin{aligned} L(\sigma_n \mid [0, s_n]) &\geq d(q, p) - d(q_n, p) + L(\sigma_n \mid [a_n, b_n]) + d(p, \gamma(r_n)) - \delta_n - \epsilon_n \\ &= d(q, \gamma(r_n)) - d(q_n, p) + L(\sigma_n \mid [a_n, b_n]) - \delta_n - \epsilon_n. \end{aligned}$$

By Lemma 14.5–(6), we have $d(q_n, p) \to 0$ as $n \to +\infty$. Also by construction, employing the normal neighborhood U, we have $L(\sigma_n \mid [a_n, b_n]) \to 0$ as $n \to +\infty$. Hence $\{\sigma_n\}$ is limit maximizing as desired.

By Lemma 14.8, any limit curve σ of this sequence is a maximal nonspacelike ray issuing from $\sigma(0) = q$. From the construction of the σ_n's, the ray σ must consist of the portion of γ from q to p, followed by α. But then by the usual "rounding the corner" technique, σ cannot be maximal unless it is unbroken, which implies that α is contained in γ as desired. □

Recalling that the generalized timelike co-ray condition is an open condition, the following consequence is obtained from Proposition 14.21.

Corollary 14.22. *Let (M, g) be timelike geodesically complete, and let $\gamma : [0, +\infty) \to (M, g)$ be a future directed timelike ray issuing from $p = \gamma(0)$ in M. Then the generalized timelike co-ray condition (with $S = \{\gamma(0)\}$) holds on an open set containing $\gamma - \{p\}$.*

In order to obtain the *smoothness* for large s of the (continuous) upper support functions $b_{p,s}$ to the Busemann function given in equation (14.21), it is necessary to have some control of the timelike cut locus. This was done in Galloway and Horta (1995, Proposition 3.9). The situation is complicated by the assumption of timelike geodesic completeness rather than global hyperbolicity, so that the theory of Chapter 9 is not immediately applicable.

Proposition 14.23. *Let (M, g) be future timelike geodesically complete, and let γ be a timelike S-ray. Suppose that the generalized timelike co-ray condition holds at $p \in I^-(\gamma) \cap I^+(S)$, and let $\alpha : [0, +\infty) \to (M, g)$ be a timelike asymptote to γ at p. Then for each $r > 0$, there exists a neighborhood U of the maximal segment $\alpha([0, r])$ which does not meet the past timelike cut locus of $\alpha(r)$.*

Proof. By taking unions, it is sufficient to show that for each s in $[0, r]$, there exists a neighborhood U of $\alpha(s)$ which does not meet the past timelike cut locus of $\alpha(r)$. There are three cases to consider: $s = r$; $0 < s < r$; and $s = 0$.

Case (1) $s = r$. Since α is a timelike ray, strong causality holds at $\alpha(r)$ by Proposition 4.40. Hence by Theorem 4.27, there exists a convex normal neighborhood U of $\alpha(r)$ such that the global Lorentzian distance function $d(g)$ agrees with the local distance function (D, U). Thus nonspacelike geodesics emanating from the center $\alpha(r)$ of this convex neighborhood contain no nonspacelike cut points within U, and U itself provides the required neighborhood.

Case (2) $0 < s < r$. This case, which essentially corresponds to Newman (1990, Lemma 3.10), is similar to the $s = 0$ case but simpler. Thus details will be omitted.

Case (3) $s = 0$. This is the case which makes use of the generalized timelike co-ray condition. It is helpful to establish two sublemmas.

Sublemma 14.24. *Let the assumptions be as in Proposition 14.23. Then for any $r > 0$, the function $x \mapsto d_r(x) = d(x, \alpha(r))$ is continuous at p.*

Proof of Sublemma 14.24. Let $p_n \to p$. Starting with Lemma 14.5–(5) and the lower semicontinuity of distance, we have for the Busemann function $b = b_\gamma$ of γ that

$$\begin{aligned} b(\alpha(r)) - b(p_n) &\geq d(p_n, \alpha(r)) \\ &\geq d(p, \alpha(r)) - \delta_n \\ &= b(\alpha(r)) - b(p) - \delta_n, \end{aligned}$$

the last equality from Corollary 14.13. Since the generalized timelike co-ray condition is assumed to hold at p, the Busemann function b is continuous at p from Theorem 14.19. Hence $d(p_n, \alpha(r)) \to d(p, \alpha(r))$ as $n \to +\infty$. □

Sublemma 14.25. *Let the assumptions be as in Proposition 14.23. Then for each $r > 0$ there exists a neighborhood U of p such that for each q in U, there exists a maximal timelike geodesic segment from q to $\alpha(r)$.*

Proof of Sublemma 14.25. Suppose the lemma is false. Then there exist $r > 0$, $p_n \to p$, and a limit maximizing sequence of past directed causal curves

$\sigma_{n,k} : [0, a_{n,k}] \to (M, g)$ such that for each k, $\sigma_{n,k}(0) = \alpha(r)$ and $\sigma_{n,k}(a_{n,k}) = p_n$. Furthermore for each fixed n, by selecting a suitable subsequence, we may suppose that as $k \to +\infty$ we have $\sigma_{n,k} \to \sigma_n$, where σ_n is either a past inextendible null ray or a past inextendible timelike ray. As was done in Lemma 14.15, diagonalize to obtain a limit maximizing sequence $\{\eta_n : [0, a_n] \to (M, g)\}$ with $\eta_n(0) = \alpha(r)$, $\eta_n(a_n) = p_n$, $a_n \to +\infty$, and with $\{\eta_n\}$ converging to a past inextendible ray $\eta : [0, +\infty) \to (M, g)$ where $\eta(0) = \alpha(r)$. (Note here that the curves are parametrized with respect to the auxiliary Riemannian metric h and also that η plays no explicit role in the rest of the proof other than serving as a reference frame for the curves η_n.)

Fix $r_0 > r$, and consider the past directed nonspacelike curves $\widetilde{\eta}_n : [0, c_n] \to (M, g)$ obtained by concatenating each η_n with the segment along $-\alpha$ from r_0 to r and reparametrizing appropriately (here $c_n = l + a_n$ where $l = r_0 - r$). Note that $\widetilde{\eta}_n(0) = \alpha(r_0)$, $\widetilde{\eta}_n(l) = \alpha(r)$, and $\widetilde{\eta}_n(c_n) = p_n$. Also put $\delta_n = d_r(p) - d_r(p_n) \to 0$ as $n \to +\infty$ by Sublemma 14.24.

Now recalling that the original sequence $\{\eta_n\}$ is limit maximizing, the limit maximality of the new sequence $\{\widetilde{\eta}_n\}$ may be established. For

$$\begin{aligned}
L(\widetilde{\eta}_n \mid [0, c_n]) &= d(\alpha(r), \alpha(r_0)) + L(\widetilde{\eta}_n \mid [l, c_n]) \\
&\geq d(\alpha(r), \alpha(r_0)) + d(p_n, \alpha(r)) - \epsilon_n \\
&= d(\alpha(r), \alpha(r_0)) + d(p, \alpha(r)) - \delta_n - \epsilon_n \\
&= d(p, \alpha(r_0)) - \delta_n - \epsilon_n
\end{aligned}$$

where $\delta_n, \epsilon_n \to 0$.

Since $\{\widetilde{\eta}_n\}$ is limit maximizing, this sequence converges to a past directed nonspacelike geodesic ray $\widetilde{\eta}$ with $\widetilde{\eta}(0) = \alpha(r_0)$. Since each $\widetilde{\eta}_n$ contains the segment of α from $\alpha(r)$ to $\alpha(r_0)$, the limit ray must be timelike and further must coincide with α back to $\alpha(0) = p$. Hence $\widetilde{\eta}$ passes through p, and yet, by construction, is contained in $\overline{J^+(p)} = \overline{I^+(p)}$. This implies, using Proposition 3.7–(1), that p is in $I^+(p)$, which contradicts the finiteness of $d(p, p)$. □

Proof of Case (3) of Proposition 14.23: $s = r$. Let all timelike geodesics, including α, be parametrized with respect to Lorentzian arc length. Suppose the proposition fails in this case. Then there exist $r > 0$ and $p_n \to p$ such that

p_n is a past timelike cut point to $\alpha(r)$ for all n. Recall from Section 9.1 that this means that there exists a past directed timelike geodesic

$$c_n : [0, d(p_n, \alpha(r)) + \delta_n) \to (M, g)$$

with $c_n(0) = \alpha(r)$ and $c_n(d(p_n, \alpha(r))) = p_n$ such that $c_n \,|\, [0, d(p_n, \alpha(r))]$ is maximal but that $d(c_n(t), \alpha(r)) > L(c_n \,|\, [0, t])$ for any $t > d(p_n, \alpha(r))$. Because α is maximal, p and $\alpha(r)$ are not conjugate along α. Hence there is a neighborhood V of $-\alpha'(r)$ in $T_{\alpha(r)}M$ which is diffeomorphic under $\exp_{\alpha(r)}$ to a neighborhood U of p. By re-indexing $\{p_n\}$ and shrinking U if necessary, we may assume that $\{p_n\} \subseteq U$ and that U satisfies the conclusions of Sublemma 14.25, so that there is a maximal timelike segment from $\alpha(r)$ to each point of U.

Let c_n be a maximal segment from $\alpha(r)$ to p_n for each n as described above so that p_n is the past timelike cut point to $\alpha(r)$ along c_n. If $c_n'(0) \in V$, by travelling along c_n a little further than p_n, we obtain a point q_n in U with $q_n \ll p_n$ and such that c_n from $\alpha(r)$ to q_n is *not* maximal. By using Sublemma 14.25, we then obtain a maximal timelike segment γ_n from $\alpha(r)$ to q_n which must satisfy $\gamma_n'(0) \notin V$. In this case, replace p_n with q_n.

Hence we may assume that we have maximal timelike geodesic segments c_n from $\alpha(r)$ to p_n with $c_n'(0) \notin V$ for any n. By passing to a subsequence and reasoning as in Sublemma 14.25, we obtain a maximal geodesic segment c from $\alpha(r)$ to p. Since α is also a maximal timelike segment from $\alpha(r)$ to p, we must have that c is timelike. Also, since $c'(0)$ is a limit point of $\{c_n'(0)\}$, we must have $c'(0) \notin V$. But this then forces α and c to be two distinct maximal timelike segments from p to $\alpha(r)$. But then α from p to $\alpha(r + \delta)$ is not maximal, contradicting the fact that α was a timelike ray. □

Reexamination of the arguments employed in the proof of Proposition 14.23 shows that the following alternative result may be proved about the timelike cut locus and a maximal timelike geodesic segment [cf. Galloway and Horta (1995, Proposition 3.9)]. Let $\gamma : [0, r_0] \to (M, g)$ be a maximal timelike geodesic segment in an arbitrary space–time, and suppose that for each r with $0 < r < r_0$, the distance function

$$x \mapsto d(x, \gamma(r))$$

is finite-valued on a neighborhood of $\gamma(0)$. Then for each r with $0 < r < r_0$, there exists a neighborhood U of the segment $\gamma([0,r])$ which does not meet the past timelike cut locus of $\gamma(r)$.

Now that control of the cut locus has been established, it is standard in the above setting that the distance function $x \mapsto d_r(x) = d(x, \alpha(r))$ is not only continuous near p as established in Sublemma 14.24 but also is *smooth* near p. (This actually follows without needing continuity first since the control of the timelike cut locus ensures that Lorentzian distance may be expressed in terms of the length in the appropriate tangent space of exponential inverses of points in M [cf. Lemma 4.26].)

We turn to an estimate (14.29) for the d'Alembertian $\Box(d_r) = \operatorname{tr} \circ H^{d_r}$, which is an important ingredient in the proof of the splitting theorem. Similar estimates have been important for a long time in global Riemannian geometry [cf. Calabi (1957) for an early illustration]. In our current setting with U chosen as in the proof of case (3) of Proposition 14.23, we first note the following basic result.

Lemma 14.26. *For q in U, let $t_0 = d(q, \alpha(r))$, and let $c : [0, t_0] \to (M, g)$ be the past directed unit speed maximal timelike geodesic segment with $c(0) = \alpha(r)$ and $c(t_0) = q$. Put $f(x) = d_r(x) = d(x, \alpha(r))$. Then*

(1) $$(\operatorname{grad} f)(q) = -c'(t_0).$$

Hence,

(2) $$\langle \operatorname{grad} f, \operatorname{grad} f \rangle = -1,$$

and the past directed maximal timelike geodesic segments from $\alpha(r)$ to points of U are integral curves of $-\operatorname{grad} f$ in U.

Proof. Let $v \in T_qM$ be given, and set $c_v(s) = \exp_q(sv)$. For s sufficiently small, we may define $\gamma_s : [0, d(c_v(s), \alpha(r))] \to (M, g)$ to be the unique maximal unit speed timelike geodesic segment from $\alpha(r)$ to $c_v(s)$. Then setting $\gamma(t, s) = \gamma_s(t)$ defines a variation of the geodesic c through geodesics with $\gamma(0, s) = \alpha(r)$

for all s. Hence by Corollary 12.24,

$$\begin{aligned}\langle(\operatorname{grad} f)(q), v\rangle = v(f) &= \frac{d}{ds}\left[d(c_v(s), \alpha(r))\right]\big|_{s=0} \\ &= \frac{d}{ds}[L(\gamma_s)]\big|_{s=0} \\ &= -\langle v, c'(t_0)\rangle\end{aligned}$$

for any v. Thus, $(\operatorname{grad} f)(q) = -c'(t_0)$ is future directed timelike, and

$$\langle \operatorname{grad} f, \operatorname{grad} f\rangle = -1$$

on U. □

Lemma 14.26 then implies that we are now precisely in a well-known geometric setup which is studied in Appendix B. We do our curvature calculations in this framework. Thus let U_1 be an open subset of (M, g), and let $f : U_1 \to \mathbb{R}$ be a smooth function on U_1 which satisfies the eikonal differential equation $\langle \operatorname{grad} f, \operatorname{grad} f\rangle = -1$. By Lemma B.1, the integral curves of $\operatorname{grad} f$ are timelike geodesics. Moreover, we note in Section B.1 that the (wrong way) Cauchy–Schwarz inequality for timelike tangent vectors implies that any such integral curve of $\operatorname{grad} f$ is maximal in the space–time $(U_1, g|_{U_1})$. We also note that curvature calculations lead to the formulas (B.22) and (B.23) relating the Ricci curvature of the integral curves of f to the d'Alembertian $\Box(f) = \operatorname{tr} \circ H^f$. Helpful in these calculations is the basic identity

$$H^f(v, \operatorname{grad} f) = 0$$

for any v in $T(U_1)$, since

$$H^f(v, \operatorname{grad} f) = \langle \nabla_v \operatorname{grad} f, \operatorname{grad} f\rangle = \frac{1}{2}v(-1) = 0.$$

In view of our desired application for which $f = d_r$, we let c denote an integral curve in U_1 of $X = -\operatorname{grad} f$. (Thus there are certain sign differences with (B.22) and (B.23) in Appendix B in which c is an integral curve of $\operatorname{grad} f$.) Recall that $\nabla_X X = 0$ on U_1 and also $H^f(X, Y) = 0$ for any Y in $T(U_1)$.

Choose an orthonormal frame $\{E_1, E_2, \ldots, E_n = X\}$ along c and neighboring integral curves. Also, let V be any parallel field along c, also extended to these nearby integral curves as a vector field parallel along them as well. Then

$$[V, X] = \nabla_V X - \nabla_{c'} V = \nabla_V X$$

since V is parallel along c. Hence,

$$\begin{aligned} R(V, c')c' &= \nabla_V \nabla_X X - \nabla_{c'} \nabla_V X - \nabla_{[V,X]} X \\ &= -\nabla_{c'} \nabla_V X - \nabla_{(\nabla_V X)} X. \end{aligned}$$

Thus

$$\begin{aligned} \langle R(V, c')c', V\rangle &= -\langle \nabla_{c'} \nabla_V X, V\rangle - \langle \nabla_{(\nabla_V X)} X, V\rangle \\ &= -c' \left(\langle \nabla_V X, V\rangle\right) + \langle \nabla_V X, \nabla_{c'} V\rangle - \sum_{j=1}^{n-1} \langle \nabla_V X, E_j\rangle \langle \nabla_{E_j} X, V\rangle \end{aligned}$$

using $\nabla_V X = \sum_{j=1}^{n-1} \langle \nabla_V X, E_j\rangle E_j$, valid since $H^f(V, E_n) = 0$. Recalling that $\nabla_{c'} V = 0$ and $X = -\operatorname{grad} f$, we then obtain

$$\langle R(V, c')c', V\rangle = \frac{d}{dt}\left\{H^f(V, V) \circ c\right\} - \sum_{j=1}^{n-1} \left(H^f(V, E_j)\right)^2.$$

If we put $V = E_i$ in this last equation and sum, we obtain

$$\begin{aligned} -\operatorname{Ric}(c', c') &= -\left(\Box(f) \circ c\right)' + \sum_{i,j=1}^{n-1} \left(H^f(E_i, E_j)\right)^2 \\ &\geq -\left(\Box(f) \circ c\right)' + \sum_{i=1}^{n-1} \left(H^f(E_i, E_i)\right)^2. \end{aligned}$$

The Cauchy–Schwarz inequality yields

$$\sum_{i=1}^{n-1} \left(H^f(E_i, E_i)\right)^2 \geq \frac{1}{n-1}\left[\sum_{i=1}^{n-1} H^f(E_i, E_i)\right]^2 = \frac{1}{n-1}\left(\Box(f) \circ c\right)^2.$$

Thus we are led to the Riccati inequality

$$(14.28) \qquad \left(\Box(f) \circ c\right)' \geq \operatorname{Ric}(c', c') + \frac{1}{n-1}\left(\Box(f) \circ c\right)^2,$$

somewhat reminiscent of the Raychaudhuri equation (12.2). Of course, it has long been known in the theory of conjugate O.D.E.'s that such Riccati inequalities lead to estimates on growth of solutions. In this particular case, we will verify that *provided* $\mathrm{Ric}(c', c') \geq 0$, integration techniques like those already encountered in Proposition 12.9 may be used to obtain the desired estimate for the d'Alembertian of $f = d_r$:

$$\Box(d_r)(q) \geq -\frac{n-1}{d(q, \alpha(r))} \tag{14.29}$$

for any q in U. Hence, a key aspect of the hypothesis in the Lorentzian Splitting Theorem that $\mathrm{Ric}(v, v) \geq 0$ for all timelike v is to ensure that estimate (14.29) is valid.

To aid in the derivation of (14.29), introduce the notation $\phi(t) = \Box \circ d_r(c(t))$, where $c : [0, d(q, \alpha(r))] \to (M, g)$ denotes the past directed maximal timelike geodesic segment from $\alpha(r) = c(0)$ to $q = c(d(q, \alpha(r)))$. Put $t_0 = d(q, \alpha(r))$. Under the Ricci curvature assumption, inequality (14.28) implies

$$\phi'(t) \geq \frac{1}{n-1}(\phi(t))^2. \tag{14.30}$$

It may be checked that $\phi(t) \to -\infty$ as $t \to 0^+$ by identifying $\phi(t)$ as the trace of the second fundamental form of the "distance sphere" of $\alpha(r)$ through $c(t)$ or by checking for Minkowski space [cf. Beem, Ehrlich, Markvorsen, and Galloway (1985, p. 38)]. Now

$$\frac{d}{dt}\left[\frac{1}{\phi(t)}\right] = -\frac{\phi'(t)}{(\phi(t))^2} \leq -\frac{1}{n-1}$$

in view of (14.30). Integrating from $t = 0$ to $t = t_0$ yields

$$\int_{t=0}^{t_0} \left[\frac{1}{\phi(t)}\right]' dt \leq -\int_{t=0}^{t_0} \frac{1}{n-1}\, dt,$$

so that

$$\frac{1}{\phi(t_0)} \leq -\frac{t_0}{n-1}$$

or

$$\phi(t_0) \geq -\frac{n-1}{t_0}.$$

Hence, recalling that $t_0 = d(q, \alpha(r))$, inequality (14.29) is established.

14.3 The Level Sets of the Busemann Function

In the differential geometry of simply connected, complete Riemannian manifolds of nonpositive sectional curvature, the level sets of the Busemann function are called "horospheres" and have been much studied in differential geometry and dynamical systems. Indeed, the already existing notion of a horosphere for the Poincaré disk or upper half plane with the usual complete Riemannian metric of constant curvature -1 seems to have motivated Busemann to define what is now called the Busemann function [cf. Busemann (1932)]. Much later, in the proof of the Cheeger–Gromoll Riemannian Splitting Theorem, the level sets of the Busemann function provided the Riemannian factor in the isometric splitting. This aspect of the Busemann function suggested that perhaps the Lorentzian Busemann function should be studied as a means toward proving a Lorentzian Splitting Theorem. So far in space–time geometry, the "horospheres" have primarily been employed as a tool in proving the space–time splitting theorem.

Before returning to Galloway and Horta (1995), it is necessary to introduce a standard concept from general relativity which has not been previously encountered in this book. Let A be an *achronal* subset of a space–time, i.e., no two points of A are chronologically related.

Definition 14.27. (*Edge of an Achronal Set*) The *edge*, edge(A), of the achronal set A consists of all points p in $\overline{A}$ such that every neighborhood U of p contains a timelike curve from $I^-(p, U)$ to $I^+(p, U)$ which does *not* meet A. The set A is said to be *edgeless* if edge$(A) = \emptyset$.

Evidently, edge$(A) \subseteq \overline{A}$. O'Neill (1983, pp. 413–415) gives an excellent exposition of the basic facts concerning edge(A) and topological hypersurfaces:

(1) The "achronal identity":

$$\overline{A} - A \subseteq \text{edge}(A); \tag{14.31}$$

(2) An achronal set A is a topological hypersurface if and only if $A \cap \text{edge}(A) = \emptyset$; and
(3) An achronal set A is a closed topological hypersurface if and only if edge$(A) = \emptyset$.

Note that being edgeless is somewhat like being "locally" a Cauchy surface in that future timelike curves starting at points in $I^-(A)$ sufficiently close to A must intersect $\overline{A}$.

Galloway and Horta (1995) adopt the following definition of a topological spacelike hypersurface. Recall that an acausal set fails to contain any causally related points and is thus also achronal.

Definition 14.28. (*Spacelike Hypersurface*) A subset S of (M, g) is said to be a *C^0-spacelike hypersurface* if for each p in S, there exists a neighborhood U of p in M such that $S \cap U$ is acausal and edgeless in $(U, g|_U)$.

The basic "edge theory" previously mentioned ensures that a spacelike hypersurface according to Definition 14.28 is an embedded topological submanifold of M of codimension one. A *smooth* spacelike hypersurface, i.e., a smooth codimension one submanifold with everywhere timelike normal, is a spacelike hypersurface in the sense of Definition 14.28. A somewhat stronger concept is that of a partial Cauchy surface.

Definition 14.29. (*Partial Cauchy Surface*) A subset S of (M, g) is said to be a *partial Cauchy surface* if S is acausal and edgeless in M.

In particular, a partial Cauchy surface is a spacelike hypersurface in the above sense and is also closed.

Now we are ready to show that the level sets of the Busemann function of a timelike S-ray are locally partial Cauchy surfaces, provided the generalized timelike co-ray condition holds [cf. Galloway and Horta (1995, Proposition 4.2), Galloway (1989a, Lemma 2.3)].

Proposition 14.30. *Let (M, g) be future timelike geodesically complete, and let γ be a timelike S-ray. Suppose that $p \in I^-(\gamma) \cap I^+(S)$ is a point of a level set $\{b_\gamma = c\}$ where the generalized timelike co-ray condition holds. Then there exists a neighborhood U of p such that the set $\Sigma_c = U \cap \{b_\gamma = c\}$ is acausal in M and edgeless in U. In particular, Σ_c is a partial Cauchy surface in $(U, g|_U)$.*

Proof. Choose an open neighborhood U of p so that $b = b_\gamma$ is continuous on U by Theorem 14.19, and also such that Lemmas 14.15 and 14.16 and

Corollary 14.17 hold on U. Since $b(q) \geq b(p) + d(p,q)$ by Lemma 14.5, we have b strictly increasing along timelike curves, which implies that Σ_c is achronal. Suppose p is an edge point of Σ_c in U. Then there exist $x, y \in U - \Sigma_c$ and future timelike curves c_1, $c_2 : [0,1] \to U$ joining x to y such that c_1 passes through Σ_c at exactly one point p, but c_2 does not meet Σ_c. Since b is strictly increasing along c_1, we must have $b(x) < b(p) = c < b(y)$. But then $b \circ c_2$ is a continuous function with $b \circ c_2(0) < c < b \circ c_2(1)$. Hence there exists $t_0 \in (0,1)$ such that $b \circ c_2(t_0) = 0$. Thus $c_2(t_0)$ lies on Σ_c in contradiction.

It remains to show the acausality of Σ_c. Since Σ_c is achronal, if this level set fails to be acausal, then there exist x, y in Σ_c with a future directed maximal null geodesic segment η from x to y. By choice of U, there exists a sequence $r_n \to +\infty$ such that there are maximal timelike segments α_n from y to $\gamma(r_n)$ which converge to a maximal timelike ray α with $\alpha(0) = y$. Recall also the notation $b_r(x) = d(\gamma(0), \gamma(r)) - d(x, \gamma(r))$ and the definition $b(x) = \lim_{r \to +\infty} b_r(x)$. Now an aspect of our choice of U is that the initial tangents to timelike geodesics remain bounded away from the null cone in making the asymptotic geodesic construction as a result of the neighborhood $K_C(y)$ of Lemma 14.16 and Corollary 14.17. With this control, it follows that by cutting the corner of the broken geodesic $\eta \cup \alpha_n$ and comparing with the corner of $\eta \cup \alpha$, there exists $\epsilon > 0$ such that for all n sufficiently large,

$$b_{r_n}(y) - b_{r_n}(x) = d(x, \gamma(r_n)) - d(y, \gamma(r_n)) > \epsilon.$$

Letting $r_n \to +\infty$, we obtain $b(y) - b(x) \geq \epsilon$. But since $x, y \in \Sigma_c$, we must have $b(x) = b(y)$ which furnishes the desired contradiction. □

It is fundamental in differential geometry that Jacobi vector fields arise as variation vector fields of one-parameter families of geodesics. More generally, it may be shown that J is a Lagrange tensor along a Riemannian (or timelike) geodesic if and only if J is a variation tensor field of a normal geodesic variation of c along some hypersurface S [cf. Eschenburg and O'Sullivan (1980, p. 7), or Appendix B]. In Riemannian geometry, given a complete minimal geodesic ray $c_v(t) = \exp(tv)$ with $\langle v, v \rangle = 1$, one forms the horoball B_v associated to the geodesic ray $c_v : [0, +\infty) \to (N, g_0)$ by taking the union of distance balls

$$B_v = \bigcup_{t>0} B_t(c_v(t)) = \bigcup_{t>0} \{q \in (N, g_0) : d_0(q, c_v(t)) < t\}.$$

The horosphere H_v associated to c_v is then the boundary of the horoball B_v. If H_v turns out to be sufficiently differentiable (C^2), then there is a normal geodesic variation of c_v along H_v for which the stable Jacobi tensor field D_v is a variation tensor field [cf. Eschenburg and O'Sullivan (1980, p. 8) and Lemma 12.13 for the construction of the stable Jacobi tensor field for a timelike geodesic].

In the proofs of various versions of the Lorentzian Splitting Theorem, mostly in the globally hyperbolic case (and also in separate investigations of hypersurface families in Riemannian manifolds), various estimates have been made on the Hessian of the distance function $x \mapsto d(x, \alpha(t))$ as $t \to +\infty$ where α is a timelike asymptote to γ, en route to obtaining various maximum principles for the Busemann function and its sublevel sets [cf. Eschenburg (1987, 1988, 1989), Galloway (1989a), Newman (1990)]. The discussion in the previous paragraph shows that this step in the proof of the splitting theorem corresponds to forming the horoball in Riemannian geometry. In particular, Riccati comparison techniques to estimate these Hessians have been well explored for Riemannian hypersurfaces by Eschenburg (1987, 1989), and it is noted that these techniques apply also to families of spacelike hypersurfaces in space–times. Eschenburg, Karcher (1989), and others take the viewpoint that studying parallel families of hypersurfaces can be most efficiently done by considering such families as level sets, at least locally, of a smooth function f satisfying the eikonal equation $\langle \operatorname{grad} f, \operatorname{grad} f \rangle = -1$ and studying the Riccati equation associated to the Hessian H^f or to its trace, the mean curvature, rather than by using Jacobi equation techniques. In Ehrlich and Kim (1994) and also in Appendix B, this approach is discussed in particular for the Lorentzian case with $\langle \operatorname{grad} f, \operatorname{grad} f \rangle = -1$, and in equations (B.18)–(B.21), the connection with the Raychaudhuri equation treated in Chapter 12 is made [cf. also Eschenburg (1975) for a thorough exploration of this methodology in the Riemannian setting].

We return to the case of interest in proving the Lorentzian Splitting Theorem and thus implicitly to the "horoballs" of a timelike geodesic asymptote to γ as given in Galloway and Horta (1995, Lemma 4.3 and Proposition 4.4) under the assumption of future timelike geodesic completeness rather than global hyperbolicity. In this result, all timelike geodesics are parametrized as Lorentzian unit speed curves, not as unit speed curves in the auxiliary Riemannian metric h.

Lemma 14.31. *Let (M, g) be future timelike geodesically complete, and let γ be a timelike S-ray. Assume that the generalized timelike co-ray condition holds at $p \in I^-(\gamma) \cap I^+(S)$. Then there exist a neighborhood U of p and constants $t_0 > 0$ and $A > 0$ such that for each q in U and each timelike asymptote α to γ from q we have*

$$H^{d_t}(v, v) \geq -A\,\langle v^\perp, v^\perp \rangle \tag{14.32}$$

for all $v \in T_qM$ and $t \geq t_0$, where $d_t = d(\,\cdot\,, \alpha(t))$ and $v^\perp$ is the projection onto the normal space $(\alpha'(0))^\perp$.

Proof. Let U be a neighborhood of p with compact closure on which the generalized timelike co-ray condition holds and also Lemmas 14.15 and 14.16 and Corollary 14.17 hold. By Corollary 14.17, the set of all initial tangents to asymptotes to γ starting from points of U is contained in a compact subset of the unit timelike tangent bundle. Then, by taking U and t_0 sufficiently small, we may ensure that the set $\{\alpha'(t) : 0 \leq t \leq t_0\}$ of all tangents to the initial segments of length t_0 of all asymptotes α emanating from points of U is contained in a compact subset of the unit timelike tangent bundle. Hence the set of all timelike two-planes Π containing these initial tangents is contained in a compact subset of the set of all timelike two-planes in $G_2(M)$. Thus, the set of all sectional curvatures of such planes is bounded above by some constant k; $K(\Pi) \leq k$.

Now consider $B = -H^{d_{t_0}}$ along the segment $\alpha\,|\,[0, t_0]$. $B = B_s$ corresponds to the second fundamental form of the level sets

$$\{q \in I^-(\alpha(t_0)) : d(q, \alpha(t_0)) = s\}$$

at $\alpha(t_0 - s)$ for $0 \leq s < t_0$. Moreover, $u \mapsto B_u$, $u = t_0 - s$, obeys the Riccati equation ${B_u}' + {B_u}^2 + R(\,\cdot\,, \alpha')\alpha' = 0$ [cf. Eschenburg (1987, Equation (3))].

Hence given the sectional curvature bound $K(\Pi) \leq k$, the appropriate Riccati Comparison Theorem may be applied [e.g., Proposition 2.4 in Eschenburg (1987)] to deduce the existence of a constant $A = A(k)$ such that for all v in T_qM

$$H^{d_{t_0}}(v, v) \geq -A\,\langle v^\perp, v^\perp \rangle.$$

The lemma now follows since $t \mapsto H^{d_t}(v, v)$ is an increasing function of t [cf. the proof of Lemma 12.13]. □

Before proving the next result and obtaining two corollaries, which are employed in the proof of the Lorentzian Splitting Theorem, we recall from the Preface of Protter and Weinberger (1984, p. v) a general philosophy important in P.D.E.'s of which Proposition 14.32 is an example.

> One of the most useful and best known tools employed in the study of partial differential equations is the maximum principle. This principle is a generalization of the elementary fact of calculus that any function $f(x)$ which satisfies the inequality $f'' > 0$ on an interval $[a, b]$ achieves its maximum value at one of the endpoints of the interval. We say that solutions of the inequality $f'' > 0$ satisfy a *maximum principle*. More generally, functions which satisfy a differential inequality in a domain D and, because of it, achieve their maxima on the boundary of D are said to possess a maximum principle.

We now turn to the proof of Proposition 14.32 following Galloway and Horta (1995, Proposition 4.4). An earlier result for globally hyperbolic space–times was obtained in Galloway (1989a, Lemma 2.4), where the opposite convention for choice of normal was used for calculation of the mean curvature [cf. Newman (1990, Lemma 3.14)].

Proposition 14.32. *Let (M, g) be future timelike geodesically complete and suppose also that $\mathrm{Ric}(v, v) \geq 0$ for all timelike v. Let γ be a timelike S-ray, and let $W \subseteq I^-(\gamma) \cap I^+(S)$ be an open set on which the generalized timelike co-ray condition holds. Let $\Sigma \subseteq W$ be a connected smooth spacelike hypersurface with nonpositive mean curvature, $H_\Sigma = \mathrm{div}_\Sigma(N) \leq 0$, where N is the future pointing unit normal along Σ. If the Busemann function $b = b_\gamma$ attains a minimum along Σ, then b_γ is constant along Σ.*

Proof. Suppose that b achieves a minimum $b(q) = a$ at q in Σ. Let U be a neighborhood of q on which the generalized timelike co-ray condition holds and also Lemmas 14.15, 14.16, and 14.31 hold. Since $b \,|\, \Sigma$ is continuous, it suffices to show that $b = a$ in a neighborhood of q in Σ. If this is not correct, then there exists a coordinate ball B with $\overline{B} \subseteq \Sigma \cap U$ centered at q such that $\partial(B) \neq \partial^0(B)$, where

$$\partial^0(B) = \{x \in \partial(B) : b(x) = a\}.$$

Note that $b > a$ on $\partial(B) - \partial^0(B)$. Also it follows from Lemma 14.31 that there exists a constant $C > 0$ such that

$$H^{d_t}(v, v) \geq -C \tag{14.33}$$

for all asymptotes at x, for all x in B, for all $v \in T_x\Sigma$ with $\langle v, v \rangle \leq 1$, and for all t sufficiently large.

By choosing B sufficiently small, we may construct a smooth function h on Σ having the following properties [cf. Eschenburg and Heintze (1984), Newman (1990, p. 177)]:

(1) $h(q) = 0$;
(2) $\|\nabla_\Sigma h\| \leq 1$ on B, where ∇_Σ denotes the gradient operator on Σ;
(3) $\Delta_\Sigma h \leq -D$ on B, where D is a positive constant and Δ_Σ is the induced Laplacian on Σ; and
(4) $h > 0$ on $\partial^0(B)$.

Consider the function $f_\epsilon = b + \epsilon h$. Note that $f_\epsilon(q) = a$, and for ϵ sufficiently small, $f_\epsilon(x) > a$ for all x in $\partial(B)$. Thus f_ϵ attains a minimum on B, say at $p = p(\epsilon)$.

Let $\alpha : [0, +\infty) \to (M, g)$ be a timelike asymptote to γ with $\alpha(0) = p$. Then for each $t > 0$, recalling Lemma 14.12 and Proposition 14.23,

$$b_{p,t}(x) = b(p) + t - d(x, \alpha(t)) \tag{14.34}$$

is a smooth upper support function for b at p. Thus the function

$$f_{\epsilon,t} = b_{p,t} + \epsilon h$$

is a smooth upper support function for f_ϵ at p, which implies that $f_{\epsilon,t}$ also has a minimum at p. We will obtain a contradiction by calculating $\Delta_\Sigma f_{\epsilon,t}(p)$ and showing that it is negative for ϵ sufficiently small and t sufficiently large.

First we digress to give a general calculation for a smooth function $f : M \to \mathbb{R}$ relating the d'Alembertian $\Box(f)$ of f to the Laplacian $\Delta_\Sigma(f \,|\, \Sigma)$ of f restricted to the spacelike hypersurface Σ. Fix $p \in \Sigma$, and let $\{e_1, e_2, \ldots, e_{n-1}\}$ be an orthonormal basis of spacelike vectors for $T_p\Sigma$. Combining the sign conventions of Galloway and Horta (1995) with those in Chapter 3, Definition 3.48 for the second fundamental form, we may define the *mean curvature* $H_\Sigma(p)$ of Σ at p as

$$H_\Sigma(p) = -\sum_{i=1}^{n-1} S_N(e_i, e_i) = \sum_{i=1}^{n-1} \langle \nabla_{e_i} N, e_i \rangle. \tag{14.35}$$

For the purposes of the calculations of Sublemma 14.33, we denote the gradient in (M, g) by $\overline{\nabla}(f)$ and the gradient in $(\Sigma, g\,|_\Sigma)$ by $\nabla(f)$. Also, let $\overline{\nabla}$ denote the Levi–Civita connection for (M, g), let ∇ denote the Levi–Civita connection for $(\Sigma, g\,|_\Sigma)$, and let H_Σ^f denote the hessian of $f \,|\, \Sigma$ in $(\Sigma, g\,|_\Sigma)$.

Sublemma 14.33. *Let Σ be a spacelike hypersurface of (M, g), and let $f : M \to \mathbb{R}$ be a smooth function. Then for any p in Σ and with the mean curvature $H_\Sigma(p)$ given by (14.35), we have*

(1) $\nabla(f \,|\, \Sigma) = \overline{\nabla}(f) + N(f)N$, *and*
(2) $\Delta_\Sigma(f \,|\, \Sigma)(p) = \Box(f)(p) + H^f(N, N)\big|_p + \langle \overline{\nabla} f, N \rangle\big|_p \, H_\Sigma(p)$.

Proof of Sublemma. (1) We have

$$\begin{aligned}\overline{\nabla}(f)(p) &= \sum_{i=1}^{n-1}\langle\overline{\nabla}(f)(p), e_i\rangle e_i - \langle\overline{\nabla}(f)(p), N(p)\rangle N(p) \\ &= \sum_{i=1}^{n-1} e_i(f)e_i - N(f)N(p) \\ &= \nabla(f \mid \Sigma) - N(f)N(p).\end{aligned}$$

(2) Now

$$\begin{aligned}\Box(f)(p) &= \sum_{i=1}^{n-1} g(e_i, e_i)H^f(e_i, e_i) + g(N(p), N(p))\, H^f(N(p), N(p)) \\ &= \sum_{i=1}^{n-1} H^f(e_i, e_i) - H^f(N, N)\big|_p.\end{aligned}$$

Hence we must calculate $H^f(e_i, e_i)$ using (1). Now

$$\begin{aligned}H^f(e_i, e_i) = \langle\overline{\nabla}_{e_i}\overline{\nabla} f, e_i\rangle &= \langle\overline{\nabla}_{e_i}\nabla(f \mid \Sigma), e_i\rangle - \langle\overline{\nabla}_{e_i}N(f)N, e_i\rangle \\ &= \langle\nabla_{e_i}\nabla(f \mid \Sigma), e_i\rangle - \langle e_i(N(f))N(p) + N_p(f)\overline{\nabla}_{e_i}N, e_i\rangle \\ &= H^f_\Sigma(e_i, e_i) - N_p(f)\,\langle\overline{\nabla}_{e_i}N, e_i\rangle.\end{aligned}$$

Thus,

$$\begin{aligned}\Box(f)(p) &= \Delta_\Sigma(f \mid \Sigma)(p) - N_p(f)\sum_{i=1}^{n-1}\langle\overline{\nabla}_{e_i}N, e_i\rangle - H^f(N, N)\big|_p \\ &= \Delta_\Sigma(f \mid \Sigma)(p) - N_p(f)H_\Sigma(p) - H^f(N, N)\big|_p. \quad \Box\end{aligned}$$

We now return to the proof of Proposition 14.32. By definition of $f_{\epsilon,t}$ we have

$$\begin{aligned}\Delta_\Sigma(f_{\epsilon,t})(p) &= \Delta_\Sigma(b_{p,t})(p) + \epsilon\Delta_\Sigma h(p) \\ &\leq \Delta_\Sigma(b_{p,t})(p) - D\epsilon.\end{aligned}$$

By Sublemma 14.33 we have

$$\Delta_\Sigma(b_{p,t})(p) = \Box(b_{p,t})(p) + H_\Sigma(p)\langle\nabla b_{p,t}(p), N(p)\rangle + H^{b_{p,t}}(N(p), N(p)).$$

Recalling inequality (14.29), we have

$$\Box(b_{p,t})(p) = -\Box(d_t)(p) \leq \frac{n-1}{t}. \tag{14.36}$$

Because $f_{\epsilon,t}$ has a minimum on Σ at p, we have $\nabla_\Sigma(f_{\epsilon,t})(p) = 0$. Also, $\nabla(b_{t,p})(p) = -\alpha'(0)$ [cf. Lemma 14.26]. Now using Sublemma 14.33–(1),

$$\begin{aligned}\nabla_\Sigma(f_{\epsilon,t})(p) &= \nabla_\Sigma(b_{p,t})(p) + \epsilon\nabla_\Sigma h(p)\\ &= \nabla(b_{p,t})(p) + \langle N(p), \nabla(b_{p,t})(p)\rangle\, N(p) + \epsilon\nabla_\Sigma h(p)\\ &= -\alpha'(0) - \langle N(p), \alpha'(0)\rangle\, N(p) + \epsilon\nabla_\Sigma h(p).\end{aligned}$$

Thus, $\nabla_\Sigma(f_{\epsilon,t})(p) = 0$ implies that

$$N(p) = \langle N(p), \alpha'(0)\rangle^{-1}[-\alpha'(0) + \epsilon(\nabla_\Sigma h)(p)].$$

Since $H^{d_t}(\nabla d_t, \cdot\,) = 0$, we also have $H^{d_t}(\alpha'(0), \cdot\,) = 0$. Thus we obtain

$$H^{b_{p,t}}(N(p), N(p)) = \epsilon^2\langle N(p), \alpha'(0)\rangle^{-2} H^{b_{p,t}}(\nabla_\Sigma h(p), \nabla_\Sigma h(p)).$$

By the (wrong way) Cauchy–Schwarz inequality,

$$|\langle N(p), \alpha'(0)\rangle| \geq \|N(p)\| \cdot \|\alpha'(0)\| = 1.$$

Since $\nabla_\Sigma h(p)$ is tangent to Σ at p and also $\|\nabla_\Sigma h(p)\| \leq 1$, we may apply inequality (14.33) to deduce $H^{b_{p,t}}(\nabla_\Sigma h(p), \nabla_\Sigma h(p)) \leq C$ and hence

$$H^{b_{p,t}}(N(p), N(p)) \leq C\epsilon^2.$$

Thus under the mean curvature assumption $H_\Sigma(p) \leq 0$ and with the choice of $N(p)$ as future directed timelike, we obtain

$$\Delta_\Sigma(b_{t,p})(p) \leq \frac{n-1}{t} + C\epsilon^2$$

which then yields

$$\Delta_\Sigma(f_{\epsilon,t})(p) \leq \frac{n-1}{t} + C\epsilon^2 - D\epsilon.$$

Then taking $t > 0$ sufficiently large and $\epsilon > 0$ sufficiently small, we obtain

$$\Delta_{\Sigma}(f_{\epsilon,t})(p) < 0,$$

which is the desired contradiction. $\square$

Proposition 14.32 has the following two corollaries [cf. Galloway and Horta (1995)], which will be used in the proof of the splitting theorem in Section 14.4. Inequality (14.36) and these corollaries will indicate that the superlevel sets

$$\{q \in M : b_{\gamma}(q) \geq c\}$$

are in some sense mean convex. This has been studied in the Riemannian case in Eschenburg (1989).

Corollary 14.34. *Let (M, g) be a future timelike geodesically complete space–time which obeys the timelike convergence condition* $\mathrm{Ric}(v, v) \geq 0$ *for all timelike v, and let γ be a timelike S-ray. Let $W \subseteq I^{-}(\gamma) \cap I^{+}(S)$ be an open set on which the generalized timelike co-ray condition holds. Let Σ be a connected acausal smooth spacelike hypersurface in W with nonpositive mean curvature. Suppose that Σ and $\Sigma_c = \{b_{\gamma} = c\} \cap W$ have a point in common and that $\Sigma \subseteq J^{+}(\Sigma_c, W)$. Then*

$$\Sigma \subseteq \Sigma_c.$$

Corollary 14.35. *Let (M, g) be a future timelike geodesically complete space–time which satisfies the timelike convergence condition* $\mathrm{Ric}(v, v) \geq 0$ *for all timelike v, and let γ be a timelike S-ray. Let $W \subseteq I^{-}(\gamma) \cap I^{+}(S)$ be an open set on which the generalized timelike co-ray condition holds. Let Σ be a smooth spacelike hypersurface with nonpositive mean curvature whose closure is contained in W. Assume that Σ is acausal in W and $\overline{\Sigma}$ is compact. If* $\mathrm{edge}(\Sigma) \subseteq \{b_{\gamma} \geq c\}$, *then*

$$\Sigma \subseteq \{b_{\gamma} \geq c\}.$$

In connection with applications of their results to the rigidity question (cf. Section 14.5 of this chapter), Galloway and Horta (1995, pp. 18–20) indicate how the regularity results previously given for the Busemann function of a timelike S-ray on the set $I^-(\gamma) \cap I^+(S)$ may be extended to the larger region $I^-(\gamma) \cap [J^+(S) \cup D^-(S)]$, where $D^-(S)$ denotes the past Cauchy development (or past domain of dependence) of S, i.e.,

$$D^-(S) = \{q \in M : \text{every future inextendible nonspacelike curve from } q \text{ intersects } S\}.$$

In particular, it is noted that if S is a compact acausal spacelike hypersurface in the future timelike geodesically complete space–time (M, g), then the generalized timelike co-ray condition holds on $I^-(\gamma) \cap [J^+(S) \cup D^-(S)]$. As a consequence of these considerations, the following is obtained in Galloway and Horta (1995, Corollary 5.6).

Corollary 14.36. *Let (M, g) be a globally hyperbolic, future timelike geodesically complete space–time which contains a compact spacelike hypersurface S, and let γ be any S-ray. Then $b_\gamma : (M, g) \to (0, +\infty]$ is continuous. Moreover, the level sets $\Sigma_c = \{q \in M : b_\gamma(q) = c\}$ are partial Cauchy surfaces in (M, g).*

Galloway and Horta (1995) note that the de Sitter space–time provides an example of (M, g) which is globally hyperbolic, geodesically complete, with compact Cauchy surfaces S, yet contains S-rays γ for which $I^-(\gamma) \neq M$. Hence, the Busemann function b_γ of such a ray γ is *not* finite-valued on all of M.

14.4 The Proof of the Lorentzian Splitting Theorem

The results have now been assembled which enable us to present a proof of the Lorentzian Splitting Theorem in the timelike geodesically complete case, as originally formulated by S. T. Yau (1982) in problem number 155. A proof under this hypothesis, rather than that of global hyperbolicity, was first given in Newman (1990). The proof given in this section follows Galloway and Horta (1995), in which simplifications are obtained in the previous proofs cited in the introduction to this chapter based on the systematic use of the generalized timelike co-ray condition.

Theorem 14.37. *Let (M, g) be a timelike geodesically complete space–time which satisfies the timelike convergence condition* $\mathrm{Ric}(v, v) \geq 0$ *for all timelike vectors v. If (M, g) contains a timelike line, then (M, g) is isometric to $(\mathbb{R} \times S, -dt^2 \oplus h)$ where (S, h) is a complete Riemannian manifold.*

Earlier work on the Lorentzian Splitting Theorem involving the Busemann function approach assumed global hyperbolicity rather than timelike geodesic completeness, in part, because of the automatic existence of maximal timelike segments connecting any pair of chronologically related points. Additionally, the Lorentzian distance function of any globally hyperbolic space–time is always continuous as well as finite-valued. This version of the splitting theorem is as follows:

Theorem 14.38. *Let (M, g) be a globally hyperbolic space–time which satisfies the timelike convergence condition* $\mathrm{Ric}(v, v) \geq 0$ *for all timelike vectors v. If (M, g) contains a* complete *timelike line, then (M, g) splits as in Theorem 14.37.*

In this second version, while completeness for all timelike geodesics is not assumed, timelike completeness of *some* maximal timelike geodesic is required.

Proof of Theorem 14.37. Let $\gamma : \mathbb{R} \to (M, g)$ be any complete timelike line parametrized with $\langle \gamma'(t), \gamma'(t) \rangle = -1$. Let $-\gamma(t) = \gamma(-t)$ denote γ with the opposite orientation. As in all the proofs of the various versions of the splitting theorems, define

$$b_r^+(x) = r - d(x, \gamma(r)), \qquad b_r^-(x) = r - d(-\gamma(r), x) = r - d(\gamma(-r), x)$$

and

$$b^+ = \lim_{r \to +\infty} b_r^+ \quad \text{and} \quad b^- = \lim_{r \to +\infty} b_r^- .$$

Put $I(\gamma) = I^+(\gamma) \cap I^-(\gamma)$. Since $\gamma \,|\, [a, +\infty)$ and $-\gamma \,|\, [a, +\infty)$ are timelike rays for all $a \in \mathbb{R}$, b^+ and b^- are defined and finite-valued on $I(\gamma)$. Moreover, all the regularity results (and their time duals) hold on a neighborhood of each point of γ. It will be helpful for use in the sequel to note that Lemma 14.5–(5) translates into the following: given any p, q in $I(\gamma)$ with $p \leqslant q$, then

$$b^+(q) \geq b^+(p) + d(p, q) \tag{14.37a}$$

and

$$b^-(p) \geq b^-(q) + d(p, q). \tag{14.37b}$$

Now fix q in $I(\gamma)$. Then there exist s, $t > 0$ so that $\gamma(-s) \ll q \ll \gamma(t)$. Hence, using the reverse triangle inequality,

$$t + s = d(\gamma(-s), \gamma(t)) \geq d(\gamma(-s), q) + d(q, \gamma(t))$$

whence

$$b_t^+(q) + b_s^-(q) = [t - d(q, \gamma(t))] + [s - d(\gamma(-s), q))] \geq 0.$$

Thus $B = b^+ + b^- \geq 0$ on $I(\gamma)$. Similar calculations show that $B(\gamma(t)) = 0$ along the geodesic $\gamma(t)$ itself.

Let U be a convex normal neighborhood of $\gamma(0)$ such that all the preceding regularity properties hold for b^+ and b^- on U. Consider the "horospheres" $S^+ = \{b^+(q) = 0\} \cap U$ and $S^- = \{b^-(q) = 0\} \cap U$. By Proposition 14.30, both S^+ and S^- are partial Cauchy surfaces in U, each containing $\gamma(0)$. Using the topological manifold structure of S^+, let W be a small coordinate ball in S^+ centered at $\gamma(0)$ whose closure is also in S^+. By applying the fundamental existence result of Bartnik (1988a, Theorem 4.1) for spacelike hypersurfaces with rough boundary data, we obtain a smooth maximal (i.e., $H_\Sigma = 0$) spacelike hypersurface $\Sigma \subseteq U$ such that Σ is acausal in U with compact closure, $\text{edge}(\Sigma) = \text{edge}(W) \subseteq S^+ = \{b^+(q) = 0\} \cap U$, and Σ meets γ. By deleting points if necessary, we will also take Σ to be defined such that

$$\Sigma \cap \text{edge}(\Sigma) \neq \emptyset.$$

First apply Corollary 14.35 to conclude that $\Sigma \subseteq \{b^+(q) \geq 0\} \cap U$. But also, since $b^-(q) \geq -b^+(q)$ on $I(\gamma)$, we have simultaneously that

$$\text{edge}(\Sigma) \subseteq \{b^-(q) \geq 0\} \cap U.$$

Hence, by the time dual of Corollary 14.35 applied to $-\gamma$, we obtain

$$\Sigma \subseteq \{b^-(q) \geq 0\} \cap U.$$

Thus

$$\Sigma \subseteq \{b^+(q) \geq 0\} \cap \{b^-(q) \geq 0\}.$$

Now since $\Sigma \subseteq \{b^+(q) \geq 0\}$ and points of edge(Σ) satisfy $b^+(q) = 0$, we obtain from Corollary 14.34 that $b^+ = 0$ along Σ. Hence, if we let $x = \gamma(t_0)$ denote the point of intersection of Σ and γ, we have $b^+(x) = 0$. Thus as $B = 0$ on γ, we obtain $b^-(x) = 0$ as well. This fact and $\Sigma \subseteq \{b^-(q) \geq 0\}$ enable the time dual of Corollary 14.34 to be applied to conclude that $\Sigma \subseteq S^-$, whence

$$b^+ = b^- = 0 \quad \text{along } \Sigma. \tag{14.38}$$

Given any x in Σ, let $\alpha_x^+ : [0, +\infty) \to (M, g)$ denote any timelike asymptote to $\gamma \,|\, [0, +\infty)$ issuing from x, and let $\alpha_x^- : [0, +\infty) \to (M, g)$ denote any timelike asymptote to $\gamma \,|\, [0, -\infty) = -\gamma \,|\, [0, +\infty)$ issuing from x. With (14.38) in hand, it is now possible to establish that at any x in Σ, the asymptotic geodesics α_x^- and α_x^+ fit together at x to form a smooth maximal geodesic line, i.e., $[\alpha_x^+]'(0) = -[\alpha_x^-]'(0)$, and further, that through each such x in Σ there is exactly one timelike asymptote α_x to γ passing through x. Toward this end consider the (possibly) broken (at $t = 0$) geodesic

$$\alpha_x(t) = \begin{cases} \alpha_x^+(t) & \text{for } \quad t \geq 0, \\ \alpha_x^-(-t) & \text{for } \quad t < 0. \end{cases}$$

First, from Corollary 14.13 we obtain for $t \geq 0$ that

$$b^+(\alpha_x(t)) = b^+(x) + d(x, \alpha_x(t)) = 0 + t = t \tag{14.39}$$

and similarly $b^-(\alpha_x(t)) = -t$ for $t < 0$. Now take $t < 0$, and consider (14.37a) with $p = \alpha_x(t) = \alpha_x^-(-t)$ and $q = x$. We obtain

$$\begin{aligned} 0 = b^+(x) &\geq b^+(\alpha_x(t)) + d(\alpha_x^-(-t), x) \\ &= b^+(\alpha_x(t)) + (-t), \end{aligned}$$

whence $b^+(\alpha_x(t)) \leq t$ for all $t < 0$. On the other hand, for any $t < 0$ we have

$$\begin{aligned} B(\alpha_x(t)) &= b^+(\alpha_x(t)) + b^-(\alpha_x(t)) \\ &= b^+(\alpha_x(t)) + b^-(\alpha_x^-(-t)) \\ &= b^+(\alpha_x(t)) - t \end{aligned}$$

so that as $B \geq 0$ on $I(\gamma)$, $b^+(\alpha_x(t)) \geq t$ for all $t < 0$. Combining these two calculations yields $b^+(\alpha_x(t)) = t$ for any $t < 0$, hence for all t in $\mathbb{R}$, recalling (14.39). An analogous argument using inequality (14.37b) in place of (14.37a) shows that $b^-(\alpha_x(t)) = -t$ for any t in $\mathbb{R}$.

We are now ready to show that $\alpha_x(t)$ is globally maximal. For take any t_1, t_2 with $t_1 < 0 < t_2$. Applying (14.37a) with $p = \alpha_x(t_1)$ and $q = \alpha_x(t_2)$ gives

$$b^+(\alpha_x(t_2)) \geq b^+(\alpha_x(t_1)) + d(\alpha_x(t_1), \alpha_x(t_2))$$

whence

$$d(\alpha_x(t_1), \alpha_x(t_2)) \leq t_2 - t_1.$$

Automatically,

$$d(\alpha_x(t_1), \alpha_x(t_2)) \geq L(\alpha_x \,|\, [t_1, t_2]) = t_2 - t_1.$$

Hence, $d(\alpha_x(t_1), \alpha_x(t_2)) = t_2 - t_1$ for any $t_1 < 0 < t_2$, which implies that α_x is globally maximal. In addition, note that if α_x *were* broken at $t = 0$, then the usual "cutting the corner" argument applied at $x = \alpha_x(0)$ would produce a longer curve from $\alpha_x(-t_0)$ to $\alpha_x(t_0)$ than $\alpha_x \,|\, [-t_0, t_0]$ for a small $t_0 > 0$, contradicting the maximality of α_x just established. Thus we have seen that for x in Σ, any future directed asymptote σ_1 to $\gamma \,|\, [0, +\infty)$ with $\sigma_1(0) = x$ and any past directed asymptote σ_2 to $-\gamma \,|\, [0, +\infty)$ with $\sigma_2(0) = x$ fit together at x to form a smooth, maximal timelike geodesic line. This is of course the geometric basis for the splitting theorem to be valid in a tubular neighborhood of γ, as we now establish.

First we establish that both α_x^+ and α_x^- are Σ-rays for any x in Σ. Recall that for any q in $I^-(\gamma \,|\, [0, +\infty)) \cap I^+(\Sigma)$, we have $0 < d(\Sigma, q) \leq b^+(q)$ [cf. inequality (14.18)]. Taking $q = \alpha_x(t)$ with $t > 0$, we obtain

$$d(\Sigma, \alpha_x(t)) \leq b^+(\alpha_x(t)) = t = L(\alpha_x \,|\, [0, t]).$$

But $d(\Sigma, \alpha_x(t)) \geq d(\alpha_x(0), \alpha_x(t)) = L(\alpha_x \,|\, [0, t])$ since α_x is maximal. Hence, $d(\Sigma, \alpha_x(t)) = t = L(\alpha_x \,|\, [0, t])$ for all $t > 0$, and $\alpha_x^+ \,|\, [0, +\infty)$ is a Σ-ray as desired. Dual arguments show that α_x^- is also a Σ-ray. It is then a consequence of basic theory in the calculus of variations that α_x^+ and α_x^- are both focal point free and that α_x meets Σ orthogonally (cf. Propositions 12.25 and 12.29).

Let us now show that direct geometric arguments reveal that for any pair of distinct points p, q in Σ, $\gamma_1 = \alpha_p^+$ and $\gamma_2 = \alpha_q^+$ do not intersect. For suppose $m = \gamma_1(s) = \gamma_2(t)$ for some s, $t > 0$. First, $t = s = d(\Sigma, m)$ since both geodesics are unit speed Σ-rays. Thus put $t_0 = t = s > 0$, and note also that $L(\gamma_1 \mid [0, t_0]) = L(\gamma_2 \mid [0, t_0]) = d(\Sigma, m)$. Choose any $\delta > 0$, and let $t_1 = t_0 + \delta$ and $n = \gamma_1(t_1)$. Starting with the fact that γ_1 is a Σ-ray, we obtain

$$\begin{aligned} d(\Sigma, n) &= L(\gamma_1 \mid [0, t_1]) \\ &= L(\gamma_1 \mid [0, t_0]) + L(\gamma_1 \mid [t_0, t_1]) \\ &= L(\gamma_2 \mid [0, t_0]) + L(\gamma_1 \mid [t_0, t_1]). \end{aligned}$$

Let $c = \gamma_2|_{[0,t_0]} * \gamma_1|_{[t_0,t_1]}$, a curve from q to n which, by the previous calculation, has the property that $L(c) = d(\Sigma, n)$. On the other hand, since $p \neq q$, we have $\gamma_1'(t_0) \neq \gamma_2'(t_0)$. Hence, the usual "rounding the corner" arguments, done using a convex normal neighborhood of m, show that we may construct a future causal curve σ from q to n with $L(\sigma) > L(c)$. But then $L(\sigma) > d(\Sigma, n)$ which is impossible. Thus the future timelike geodesics α_x^+, $x \in \Sigma$, do not intersect. Dual arguments reveal that none of the past timelike geodesics α_x^-, $x \in \Sigma$, can intersect either.

Let us now turn to the non-intersection of any two geodesic rays of the form $\gamma_1 = \alpha_p^+$ and $\gamma_2 = \alpha_q^-$, for any $p, q \in \Sigma$ except possibly at $p = q$ in Σ. Suppose that there exist s, $t > 0$ such that $\gamma_1(s) = \gamma_2(t) = m$. Hence, $b^+(m) = b^+(\gamma_1(s)) = s > 0$. On the other hand, taking any t_1 with $0 < t_1 < t$, we have $b^+(\gamma_2(t_1)) = -t_1$ from our previous work in the course of this proof. Since $b^+ \circ \gamma_2$ is continuous, there exists $t_2 > 0$ such that $b^+(\gamma_2(t_2)) = 0$. But this then contradicts the achronality of the set $\{b^+ = 0\}$. Hence we have established the fact that no two normal geodesics to Σ intersect, except for the trivial possibility that both issue from the same point p of Σ and that they intersect precisely at that point or coincide.

Recall that Σ is a smooth spacelike hypersurface. Hence there exists a smooth future pointing unit timelike normal field N along Σ. Recalling also the hypothesis of timelike geodesic completeness, we may then define the normal exponential map $E : \mathbb{R} \times \Sigma \to (M, g)$ by $E(t, q) = \exp_q(tN(q))$. In view of the properties derived above for the asymptotic geodesics to γ issuing from points of Σ, we already have that E is nonsingular and injective and hence a diffeomorphism onto its image.

By our above calculations for $b^+(\alpha_x(t))$ and $b^-(\alpha_x(t))$, we have that

$$b^+(E(t, q)) = b^+(\exp(tN(q))) = t \tag{14.40}$$

and

$$b^-(E(t, q)) = b^-(\exp(tN(q))) = -t. \tag{14.41}$$

Since E is a diffeomorphism onto its image, these equations reveal that both b^+ and b^- are differentiable on the connected open set $U_1 = E(\mathbb{R} \times \Sigma)$.

Fix $x \in \Sigma$, and let $\alpha_{x,r}$ denote the unit timelike geodesic from x to $\gamma(r)$ for r sufficiently large. Then as $b_r^+(x) = r - d(x, \gamma(r)) = r - d_r(x)$, one has $\nabla b_r^+(x) = -\nabla d_r(x) = -\alpha_{x,r}'(0)$ (cf. Lemma 14.26). Hence letting $r_n \to +\infty$, and remembering that U has been chosen so that the asymptotic geodesic construction is well behaved, we have $\alpha_x{}'(0) = \lim_{n\to+\infty} \alpha_{x,r_n}'(0)$ and thus

$$\nabla b^+(x) = \lim_{n\to+\infty} \nabla b_{r_n}^+(x) = \lim_{n\to+\infty} -\alpha_{x,r_n}'(0) = -\alpha_x{}'(0),$$

where $\alpha_x(t) = E(t, x)$. More generally,

$$\nabla b^+(\alpha_x(t)) = -\alpha_x{}'(t)$$

for any $t \in \mathbb{R}$ [cf. Eschenburg (1988)]. This discussion shows that $f = b^+$ is a differentiable function on U_1 which satisfies the eikonal equation $\langle \nabla f, \nabla f \rangle = -1$. (The above fact that asymptotes to γ have tangents parallel to ∇b^+ also shows that for any point q in U_1 there exists a unique unit speed maximal timelike line asymptotic to γ passing through q.) Put

$$\Sigma_t = E(\{t\} \times \Sigma). \tag{14.42}$$

Then $N(q) = -\nabla b^+(q)$ provides a future directed unit normal to Σ_t for each q in Σ_t. As is well known [cf. Eschenburg (1975)], the expansion tensor of a certain Lagrange tensor field along any unit timelike geodesic $c(t)$ in U_1 asymptotic to γ calculates the mean curvature $H(t)$ of the leaf of this foliation at $c(t)$. The Raychaudhuri equation associated to this Lagrange tensor field then translates into the following evolution inequality [cf. Galloway and Horta (1995, p. 18)]:

$$H'(t) = -\operatorname{Ric}\big(N(c(t)), N(c(t))\big) - \|\nabla N\|^2 \leq -\frac{1}{n-1}(H(t))^2. \tag{14.43}$$

Assuming that N *fails* to be parallel along the geodesic c and recalling the curvature assumption $\operatorname{Ric}(N, N) \geq 0$, arguments similar to those in Proposition 12.9 of Chapter 12 show that the mean curvature $H(t)$ to Σ_t at $c(t)$ must blow up in finite time. But this contradicts the fact established above that all such timelike asymptotes c to γ are focal point free to Σ_0 at $c(0)$. Hence, the normal vector field $N = -\nabla b^+$ must be parallel in U_1, and the level surfaces Σ_t are all totally geodesic (cf. Lemma B.5 of Appendix B). Moreover, the parallelism of N also implies that U_1 splits locally isometrically by Wu's proof (1964, p. 299) of the Lorentzian de Rham Theorem. More exactly, let h denote the Riemannian metric induced on the spacelike hypersurface $\Sigma = \Sigma_0$ by inclusion in (M, g). Then E is an isometry of $(\mathbb{R} \times \Sigma, -dt^2 \oplus h)$ and $(U_1, g\,|_{U_1})$ and hence provides the desired isometric splitting in the neighborhood U_1 of the given timelike line.

We summarize what has been accomplished so far as

Proposition 14.39. *Suppose (M, g) is timelike geodesically complete and satisfies $\operatorname{Ric}(v, v) \geq 0$ for every timelike v. Let $\gamma : \mathbb{R} \to (M, g)$ be any complete timelike line in (M, g). Then there exists an open neighborhood U of $\gamma(\mathbb{R})$ in M such that*

(1) *$\Sigma = \{b^+ = b^- = 0\} \cap U$ is a smooth embedded acausal hypersurface of (M, g),*

(2) *$(U, g|_U)$ is isometric to $(\mathbb{R} \times \Sigma, -dt^2 \oplus h)$, where h is the metric induced on Σ by inclusion in (M, g), and*

(3) *for each q in Σ, the curve $\mathbb{R} \mapsto U$ given by $c_q(t) = \exp(tN(q))$ is a maximal timelike line passing through q which is asymptotic to the given timelike line γ. Further, any timelike asymptote to $\gamma\,|\,[a, +\infty)$ or*

$-\gamma \mid [a, +\infty)$, for any a in $\mathbb{R}$, issuing from any point q in U, coincides up to parametrization with one of these flow lines $c_q(t)$ of ∇b^+. (Here N denotes the future directed unit timelike normal to Σ.)

It is now necessary to show that the local splitting given in Proposition 14.39 may be extended to a global splitting. In Eschenburg (1988) this was carried out by defining timelike lines c_1, c_2 to be *equivalent* if they are joined by a chain of timelike lines $\beta_1 = c_1, \beta_2, \ldots, \beta_k = c_2$ such that each successive pair β_j, β_{j+1} forms the boundary of an isometrically immersed flat strip $F_j : (\mathbb{R} \times [a_1, a_2], -dt^2 \oplus ds^2) \to (M, g)$ satisfying $F_j(\mathbb{R} \times \{a_1\}) = \beta_j$, $F_j(\mathbb{R} \times \{a_2\}) = \beta_{j+1}$, and $F_j(\mathbb{R} \times \{s\})$ is a timelike line for each s in $[a_1, a_2]$. Later, following the publication of Galloway (1989a), Galloway indicated to us that an alternate proof of the global splitting could be given by maximally enlarging the local hypersurface Σ of Proposition 14.39 originally obtained in a neighborhood of a given timelike line. It is this second line of reasoning that will be pursued next.

It will be useful for this purpose to recall the following elementary relationship between parallel translation and parallel fields. Suppose that U is a (path) connected open set and N is a smooth vector field which is globally parallel on U, i.e., for any $q \in U$ and any $v \in T_qM$, we have $\nabla_v N = 0$. Fix any point q_1 in U, and let $v_1 = N(q_1) \in T_{q_1}M$. Then for any q_2 in U,

(14.44) If $c : [0, 1] \to U$ is a smooth curve with $c(0) = q_1$ and $c(1) = q_2$, and $v_2 \in T_{q_2}M$ is obtained by parallel translating v_1 along c from $t = 0$ to $t = 1$, then the resulting vector v_2 is independent of choice of path c from q_1 to q_2.

This basic fact is valid because the parallel field P along c with $P(0) = v_1$ must be given by $P(t) = N \circ c(t)$ for all t in $[0, 1]$, whence $v_2 = P(1) = N(q_2)$ which depends on N and q_2 but not on the particular choice of smooth curve c from q_1 to q_2. Thus if the field N on U is globally given, no holonomy problems are encountered in parallel translation along different curves in U.

We turn now to the problem of compatibly enlarging a local splitting of the form given in Proposition 14.39. Thus let $\gamma_0 : \mathbb{R} \to (M, g)$ denote a given timelike line, and let W be a convex normal neighborhood of $p_0 = \gamma_0(0)$ such that the regularity properties for timelike asymptote construction are valid on W and the Busemann functions b_0^+ and b_0^- are continuous on W. Let W_0 be a second convex open neighborhood of p_0 with $\overline{W}_0 \subseteq W$, which is used to obtain a local isometric splitting $E : (\mathbb{R} \times \Sigma_0, -dt^2 \oplus h_0) \to (U_0, g)$ with $\Sigma_0 \subseteq W_0$ and $E(t, q) = \exp(t\, N_0(q))$, where $N_0 = -\nabla b_0^+$ is globally parallel on $U = E(\mathbb{R} \times \Sigma_0)$. Let $p_1 \in \text{edge}(\Sigma_0)$ be arbitrary. Take $\{q_n\} \subseteq \Sigma_0$ with $\lim_{n\to+\infty} E(0, q_n) = p_1$. By continuity, $b_0^+(p_1) = \lim_{n\to+\infty} b_0^+(E(0, q_n)) = 0$, and similarly $b_0^-(p_1) = 0$. Recalling that Σ_0 is totally geodesic and W is convex, there exist unit speed h_0-geodesics $c_n : [0, a_n] \to \Sigma_0$ with $c_n(a_n) = q_n$. Let w in $T_{p_0}(\Sigma_0)$ be an accumulation vector of $\{c_n'(0)\}$ and put $c(t) = \exp(tw)$, $c : [0, a) \to (\Sigma_0, h_0)$. Since $c_2(t) = (2t, c(t))$, $0 \le t < a$, is a unit timelike geodesic in (M, g) and (M, g) is timelike geodesically complete, it follows that $c_2(t)$ extends to $t = a$. Hence, the geodesic $c : [0, a) \to (\Sigma_0, h_0)$ extends to a geodesic $c_1 : [0, a] \to (M, g)$ with $c_1(a) = p_1$.

Let P be the unique parallel field along $c_1 : [0, a] \to (M, g)$ with $P(0) = \gamma_0'(0)$. Put $v = P(1)$ in $T_{p_1}M$, and let $\gamma_1 : \mathbb{R} \to (M, g)$ be given by $\gamma_1(t) = \exp_{p_1}(tv)$. If we take $\{s_n\} \subseteq (0, a)$ with $s_n \to a$ and put $\sigma_n(t) = \exp(tP(s_n)) = \exp(tN_0(c_1(s_n)))$, then from the nature of the local splitting in Proposition 14.39, each σ_n is a maximal timelike line asymptotic to γ_0. Thus γ_1 is also a maximal timelike line.

Apply Proposition 14.39 to γ_1 to get a local splitting $E : (\mathbb{R} \times \Sigma_1, -dt^2 \oplus h) \to (U_1, g)$ in a tubular neighborhood U_1 of γ_1 with $N_1 = -\nabla b_1^+$ globally parallel on U_1, $b_1^+(p_1) = b_1^-(p_1) = 0$. We wish to show that $N_1(q) = N_0(q)$ at all points q of $U_0 \cap U_1$, or equivalently, $b_1^+(q) = b_0^+(q)$ (and thus also $b_1^-(q) = b_0^-(q)$) at all points q of $U_0 \cap U_1$.

Choose b with $0 < b < a$ so that $q = c_1(b) \in \Sigma_1 \cap \Sigma_0$, where $c_1 : [0, a] \to (M, g)$ denotes the g-geodesic in W from p_0 to p_1 constructed above. Recall that $\gamma_1'(0)$ is obtained by parallel translating $\gamma_0'(0)$ along c_1 from p_0 to p_1. Also, $N_0(q) = -\nabla b_0^+(q)$ is obtained by parallel translating $\gamma_0'(0) = N_0(p_0)$ along c_1 from $t = 0$ to $t = b$, and $N_1(q) = -\nabla b_1^+(q)$ is obtained by parallel translating $\gamma_1'(0) = N_1(p_1)$ along c_1 from $t = a$ to $t = b$. Hence by (14.44), we must have $N_1(q) = N_0(q)$. Given any other point m in $\Sigma_1 \cap \Sigma_0$, take a smooth curve $\sigma : [0,1] \to \Sigma_1 \cap \Sigma_0$ from $q = \sigma(0)$ to $m = \sigma(1)$. Since N_1 and N_0 are both parallel at all points of $\Sigma_1 \cap \Sigma_0$ and $N_1(q) = N_0(q)$, it follows again from (14.44) that $N_1(m) = N_0(m)$. Hence, $N_1 = N_0$ at all points of $\Sigma_1 \cap \Sigma_0$.

Now given any $n \in U_1 \cap U_0$, use either splitting to find a smooth curve $\mu(t) = E(t,m)$ from a point m in $\Sigma_1 \cap \Sigma_0$ to n. Since $N_1(m) = N_0(m)$ and $N_1(n)$ and $N_0(n)$ are obtained from $N_1(m)$, $N_0(m)$ by parallel translation along μ, we must have $N_1(n) = N_0(n)$.

Thus we have obtained $\nabla b_1^+ = \nabla b_0^+$ on $U_1 \cap U_0$. Since $b_1^+(q) = b_0^+(q)$ for any q in $\Sigma_1 \cap \Sigma_0$, it follows that $b_1^+ = b_0^+$ on $U_1 \cap U_0$. Put $\Sigma = \Sigma_0 \cup \Sigma_1$ furnished with the Riemannian metric h induced by inclusion in (M, g). Further, put

$$N(q) = \begin{cases} N_0(q) & \text{if } q \in \Sigma_0, \\ N_1(q) & \text{if } q \in \Sigma_1, \end{cases}$$

and define the extended splitting $E : (\mathbb{R} \times \Sigma, -dt^2 \oplus h) \to (U_0 \cup U_1, g)$ by $E(t,q) = \exp(t\,N(q))$.

Now we indicate how the original splitting may be extended from the given local splitting $\mathbb{R} \times \Sigma_0 \to U_0$ "in all directions." Select sufficiently many points $\{p_1, p_2, \ldots, p_k\}$ in edge(Σ_0) and associated convex neighborhoods $W_1, \ldots, W_k$ such that after determining local hypersurfaces $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ and tubular neighborhoods $U_1, U_2, \ldots, U_k$, that if any pair $U_p \cap U_q \neq \emptyset$ (or $\Sigma_p \cap \Sigma_q \neq \emptyset$), then $\Sigma_p \cap \Sigma_q$ contains a point of the starting hypersurface Σ_0. With this proviso, it follows as above using (14.44) that the Busemann functions (and their gradients) obtained from the local splittings $E_j : \mathbb{R} \times \Sigma_j \to U_j$ agree on all overlaps, so that we may put

$$\Sigma = \Sigma_0 \cup \Sigma_1 \cup \Sigma_2 \cup \cdots \cup \Sigma_k$$

and

$$U = U_0 \cup U_1 \cup \cdots \cup U_k$$

and define a parallel field on U by $N(q) = N_j(q)$ for q in U_j to obtain a well-defined isometric splitting $E : \mathbb{R} \times \Sigma \to (U, g\,|_U)$ via $E(t,q) = \exp(t\,N(q))$.

We now present the more general inductive step necessary to conclude that the given local splitting $E : (\mathbb{R} \times \Sigma_0, -dt^2 \oplus h) \to (U_0, g)$ may be globalized.

Thus we suppose that an open, smooth, path connected spacelike hypersurface Σ with Riemannian metric h induced by inclusion in (M, g) is given, with a timelike unit normal field N parallel at all points of Σ so that the mapping

$$E : (\mathbb{R} \times \Sigma, -dt^2 \oplus h) \to (U = E(\mathbb{R} \times \Sigma), g)$$

given by $E(t, q) = \exp_q(t\,N(q))$ is a diffeomorphism onto its image, as well as being an isometry, and that the timelike geodesics $\gamma_q(t) = E(t, q)$ are timelike lines for all q in Σ. Inductively, suppose further that if N is extended to U by setting $N(\gamma_q(t)) = \gamma_q{}'(t)$, then N is globally parallel on (U, g). Finally, although this is not necessary to the argument, suppose that Σ contains a given local hypersurface Σ_0'' formed by starting with an originally given complete timelike line γ_0 and that $N \mid U_0'' = -\nabla b^+_{\gamma_0}$, as in Proposition 14.39.

Suppose that edge(Σ) is nonempty. Take any $p \in$ edge(Σ). To complete the argument, we will show that Σ can be extended consistently past p to form an even larger spacelike hypersurface Σ_1 satisfying the above conditions. This then shows that if we want to carry out "Zorn's Lemma" type arguments, then Σ_0'' may be maximally extended to a smooth spacelike hypersurface (S, h) with edge$(S) = \emptyset$, and we may thus obtain the desired global splitting.

Let $W \subseteq U$ be a geodesically convex open set with $p \in \overline{W}$ and W contained inside a larger geodesically convex open set. By our previous discussion, if we select p_0 in W sufficiently close to p and let W_0 be a geodesically convex neighborhood of p_0 with p in $\partial(W_0)$, then applying the local splitting Proposition 14.39 produces a hypersurface Σ_0 contained in $\Sigma \cap W_0$ with $p_0 \in \Sigma_0$ and $p \in$ edge(Σ_0). Further, $N(q) = -\nabla b_0^+(q)$ at all points of Σ_0. Also, from our previous work on extending the local splitting, we may find a geodesic $c_1 : [0, 1] \to (M, g)$ with $c_1(0) = p_0$, $c_1(1) = p$, and $c_1([0, 1)) \subseteq \Sigma_0$. Then, by parallel translating $N(p_0)$ along c_1, we obtain a unit timelike vector $w_1 \in T_pM$ so that $c(t) = \exp(tw_1)$ is a complete timelike line. As in our above arguments, we now apply Proposition 14.39 to the timelike line $c(t)$ in a small convex neighborhood W_1 of $c(0)$ to obtain an open, path connected, smooth spacelike hypersurface Σ_1 containing p, and we have shown above that for all q in $U_0 \cap U_1$, we have $\nabla b_0^+(q) = \nabla b_1^+(q)$. Whence we see that taking $N = -\nabla b_1^+$, i.e., using the Busemann function associated to $c(t)$, the parallel field N may be consistently extended to $\Sigma \cup \Sigma_1$.

The crucial remaining step in this procedure is to show that this process just described is well defined. That is, suppose a different point $q_0 \in W$ is selected, the geodesic $c_2 : [0,1] \to (M,g)$ with $c_2(0) = q_0$, $c_2(1) = p$, and $c_2([0,1)) \subseteq \Sigma$ is constructed, and $w_2 \in T_pM$ is obtained by parallel translating $N(q_0)$ along c_2 from $t = 0$ to $t = 1$. It is necessary to establish that $w_1 = w_2$ in T_pM.

Thus suppose that $w_1 \neq w_2$ in T_pM. Let P_j denote the parallel field along $c_j : [0,1] \to (M,g)$ with $P_1(0) = N(p_0)$ [respectively, $P_2(0) = N(q_0)$]. Recall that in fact $P_j(t)$ is equally well determined by the value of $N(c_j(t))$ for any t with $0 < t < 1$. Thus in view of the proof of the local splitting Proposition 14.39, especially the use of Bartnik (1988a), we will replace the given initial points p_0, q_0 by points p_0' on c_1 and q_0' on c_2 sufficiently close to p such that after the associated local splittings

$$E_j : (\mathbb{R} \times \Sigma_j', -dt^2 \oplus h_j) \to (U_j', g), \qquad j = 1, 2$$

are obtained, centered about $\gamma_{p_0'}(t)$ and $\gamma_{q_0'}(t)$ respectively, we have $p \in \Sigma_1' \cap \Sigma_2'$. From our previous studies of the compatibility of local splittings, we also have $\nabla b_1^+ = \nabla b_2^+$ at all points of $\Sigma_1' \cap \Sigma_2'$, and particularly, $P_1(p_0') = -\nabla b_1^+(p_0')$ and $P_2(q_0') = -\nabla b_2^+(q_0')$. Thus, in fact, $w_1 = P_1(p) = -\nabla b_1^+(p) = -\nabla b_2^+(p) = w_2$, as desired.

Now that the existence of a global isometric splitting $(M,g) = (\mathbb{R} \times S, -dt^2 \oplus h)$ has been established, the only remaining detail in the proof is to show that the Riemannian factor (S,h) is forced to be complete by the assumed timelike completeness of (M,g). But this is the same type of argument that has already been used above. Suppose that (S,h) fails to be Riemannian complete. Then there exists an h-unit speed Riemannian geodesic $c : (a,b) \to (S,h)$ where either a or b is finite, which is inextendible either to $t = a$ or to $t = b$ or possibly both. Define $\gamma : (a,b) \to (M,g)$ by $\gamma(t) = (2t, c(t))$, which is a geodesic in (M,g) by formula (3.17) with $f = 1$. Further, $g(\gamma'(t), \gamma'(t)) = -2 + h(c'(t), c'(t)) = -1$. Thus γ is a timelike geodesic in (M,g) which is either past incomplete or future incomplete, contradicting the hypothesis of timelike geodesic completeness of (M,g). □

Remark 14.40. It is perhaps amusing to note *after the fact* that the two versions of the Lorentzian Splitting Theorem are not so far apart.

(1) Note that under the conclusion of Theorem 14.37, (M, g) must be strongly causal by Corollary 3.56. Alternatively, we know from the proof of Theorem 14.37 that a complete timelike line passes through every point of (M, g), hence strong causality also follows from Proposition 4.40.

(2) Since the Riemannian factor (S, h) in Theorem 14.37 turns out to be complete, we know that (M, g) had also to be globally hyperbolic (as well as geodesically complete) by Theorem 3.67, (1) $\Rightarrow$ (2), (3).

(3) Similarly in Theorem 14.38, the global hyperbolicity of (M, g) ensures after the splitting is obtained that (M, g) had to be geodesically complete, also by Theorem 3.67, (3) $\Rightarrow$ (2).

(4) The use of Theorem 4.1 of Bartnik (1988a) enables certain analytic arguments involving smooth super support functions and subsupport functions to show that the continuous function $b^+ \circ c$ is affine, hence differentiable, to be dispensed with in the proof presented here. An aspect of this other proof method is to observe that by estimates akin to (14.36), these support functions have arbitrarily small second derivatives [cf. Eschenburg and Heintze (1984), Beem, Ehrlich, Markvorsen, and Galloway (1985, Lemma 5.1), Eschenburg (1988, Lemma 5.2 and Proposition 6.3)].

A different application of the Busemann function methods discussed in this chapter has been given in Andersson and Howard (1994) to obtain results more closely allied to Harris's Maximal Diameter Theorem A.5 of Appendix A. In this setting, for example, estimate (14.29) translates into

$$\Box(d_r) \geq -(n-1)\tan(\pi/2 - d_r).$$

Here are two of the results obtained in Andersson and Howard (1994, Theorem 7.1, Corollary 7.5).

Theorem 14.41. *Let (M^n, g) be a globally hyperbolic space–time which satisfies the curvature condition* $\mathrm{Ric}(v, v) \geq (n-1)$ *for any unit timelike vector v. Suppose that (M, g) contains a maximal timelike geodesic segment $\gamma : (-\pi/2, \pi/2) \to (M, g)$. Then there exists a complete Riemannian manifold (S, h) such that (M, g) is isometric to the warped product*

$$\left((-\pi/2, \pi/2) \times S, -dt^2 + \cos^2(t)\, h\right).$$

As a consequence of Theorem 14.41, Andersson and Howard are able to obtain a related result to Theorem A.5 of Harris (1982b). The definition of "globally hyperbolic of order $q = 1$" is given in Appendix A, Definition A.3.

Theorem 14.42. *Let (M^n, g) be a space–time which is globally hyperbolic of order* 1 *and also satisfies the curvature condition* $\mathrm{Ric}(v, v) \geq (n-1)$ *for any unit timelike vector v.*

(1) *If (M, g) contains a maximal timelike geodesic segment $\gamma : \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to (M, g)$ of length π joining x to y, then the causal interval $D = I^+(x) \cap I^-(y)$ is isometric to the warped product of the hyperbolic model space (S, h) of constant curvature -1 and dimension $(n-1)$ and the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, with warped product metric $-dt^2 + \cos^2(t)\, h$.*
(2) *If (M, g) contains a timelike geodesic $\gamma : (-\infty, +\infty) \to (M, g)$ such that each segment $\gamma \,|\, [t, t + \pi]$ is maximal, then (M, g) is isometric to the universal cover of anti–de Sitter space.*

14.5 Rigidity of Geodesic Incompleteness

Now that the space–time splitting theorem has been obtained, we are ready to return to the question of rigidity of timelike geodesic incompleteness as discussed in the introduction to this chapter. The following formulation of this question was given in Bartnik (1988b). Bartnik defines a space–time to be *cosmological* if it is globally hyperbolic with a compact Cauchy surface and satisfies the timelike convergence condition $\mathrm{Ric}(v, v) \geq 0$ for all timelike v.

Conjecture 14.43 (Bartnik). *Let (M, g) be a cosmological space–time. Then*

(1) *(M, g) contains a constant mean curvature Cauchy surface; and*
(2) *(M, g) is either timelike geodesically incomplete or else (M, g) splits isometrically as a product $(\mathbb{R} \times V, -dt^2 \oplus h)$, where (V, h) is a complete Riemannian manifold.*

In Geroch (1966) the following condition was introduced:

(14.45) For some $p \in M$, the set $M - [I^+(p) \cap I^-(p)]$ is compact.

Bartnik (1988b) proves that if condition (14.45) is satisfied, then there is a constant mean curvature Cauchy surface in (M, g) which passes through p. Also Bartnik (1988b) notes that if condition (14.45) holds, then either (M, g) is a metric product or else is timelike geodesically incomplete, extending earlier work of Galloway (1984c) which required that this condition hold at all p on a compact Cauchy surface. However, (1) is false in general [cf. Bartnik (1988b)].

In Beem, Ehrlich, Markvorsen, and Galloway (1985), a rigidity result was stated which may be obtained as a consequence of the Lorentzian Splitting Theorem along fairly elementary lines, without any causality assumption like (14.45), but under the weaker sectional curvature hypothesis. A complete proof was published in Ehrlich and Galloway (1990).

Theorem 14.44. *Let (M, g) be a space–time with a compact Cauchy surface and everywhere nonpositive timelike sectional curvatures $K \leq 0$. Then either (M, g) is timelike geodesically incomplete or else (M, g) splits as a metric product $(\mathbb{R} \times V, -dt^2 \oplus h)$, where (V, h) is compact*

Proof. Suppose that (M, g) is timelike geodesically complete. Since (M, g) is assumed to have a compact Cauchy surface S, it is not difficult to see that (M, g) is causally disconnected with $K = S$. Hence by Theorem 8.13, (M, g) contains a nonspacelike geodesic line. If the line is shown to be timelike, then the desired splitting follows from the Lorentzian Splitting Theorem. Hence, it only remains to rule out the alternative that the line is null.

This may be accomplished by using a corollary to Harris's Toponogov Theorem of Appendix A [cf. Ehrlich and Galloway (1990)].

Theorem 14.45. *Let (M,g) be a space–time with compact Cauchy surface and with everywhere nonpositive timelike sectional curvatures. If (M,g) is future timelike geodesically complete, then each future inextendible null geodesic contains a null cut point.*

Proof. Assume that there is a future inextendible null geodesic $\eta : [0,a) \to (M,g)$ such that there is no cut point to $p = \eta(0)$ along η. Let $\{t_k\}$ be an increasing sequence of nonnegative real numbers converging to a, and set $q_k = \eta(t_k)$. Hence $d(p,q_k) = 0$ for all k since η contains no null cut point to p. Fix a compact Cauchy surface S containing p, and for each k let γ_k be a future directed timelike geodesic segment from some point r_k of S to q_k which realizes $d(S,q_k)$. (Hence γ_k is also a maximal geodesic segment.) The sequence $\{r_k\}$ has an accumulation point r on S, and thus some subsequence $\{\gamma_m\}$ of $\{\gamma_k\}$ converges to a future inextendible nonspacelike limit curve γ with $\gamma(0) = r$ which by the usual arguments is a maximal geodesic. Also, for each fixed t the segment $\gamma \mid [0,t]$ realizes the distance from S to $\gamma(t)$. Moreover, for some $t > 0$ we have $d(S,\gamma(t)) > 0$ since γ cannot be imprisoned in S. Hence, the limit geodesic must be timelike rather than null. Note also that γ is future complete by the assumption of future timelike geodesic completeness. Furthermore,

$$\gamma \cap I^+(p) = \emptyset. \tag{14.46}$$

For suppose γ meets $I^+(p)$. Then γ_m meets $I^+(p)$ for large m. Hence $q_m \in I^+(p)$, and thus $d(p,q_m) > 0$, in contradiction.

Now we use the sectional curvature hypothesis to obtain a contradiction to (14.46). Because γ is a timelike limit curve of $\{\gamma_m\}$, we may choose m sufficiently large so that $q_m \in I^+(r)$. Using $I^+(q_m) \subseteq I^+(p)$, we then obtain

$$I^+(r) \cap I^+(p) \neq \emptyset.$$

Thus we may fix x in $I^+(r) \cap I^+(p)$ and let $\delta : [0,L] \to (M,g)$ be a maximal timelike geodesic segment from r to x. Since $K \leq 0$ and γ is future complete, Theorem 3.1 of Harris (1982b, p. 313), a consequence of Harris's Toponogov Theorem, implies that there is some t_0 with $x = \delta(L) \leq \gamma(t_0)$, which yields $p \ll x \leqslant \gamma(t_0)$, whence $p \ll \gamma(t_0)$ in contradiction to (14.46). □

At the end of Section 14.3, a summary was given of how Galloway and Horta (1995, pp. 18–20) extend the regularity results for the Busemann function as given in this chapter for $I^-(\gamma) \cap I^+(S)$ to $I^-(\gamma) \cap [J^+(S) \cup D^-(S)]$. With this accomplished, Galloway and Horta (1995, Theorem 5.7) are able to improve upon earlier results in Eschenburg and Galloway (1992, Theorem B).

Theorem 14.46. *Let (M, g) be a space–time which contains a compact acausal spacelike hypersurface S and satisfies the timelike convergence condition* $\mathrm{Ric}(v, v) \geq 0$ *for all timelike v. If (M, g) is timelike geodesically complete and contains a future S-ray γ and a past S-ray η such that $I^-(\gamma) \cap I^+(\eta) \neq \emptyset$, then (M, g) splits as in the rigidity conjecture.*

An entirely unrelated investigation of rigidity of geodesic incompleteness is pursued in Garcia–Rio and Kupeli (1995). Suppose (M, g) is a stably causal space–time with smooth time function $f : M \to \mathbb{R}$, i.e., $g(\nabla(f), \nabla(f)) < 0$. Rescale the given metric g to produce a new metric g_c for M for which $g_c(\nabla^c(f), \nabla^c(f)) = -1$ by setting $g_c = -g(\nabla(f), \nabla(f)) \cdot g$. Now the integral curves of $\nabla^c(f)$ are unit speed timelike geodesics (cf. Lemma B.1). Hence, timelike geodesic completeness of (M, g_c) implies that the flow of f in (M, g_c) is complete, and the flow may be used to produce a metric $g(t)$ on the level surface $f^{-1}(t)$ from that of $\left(f^{-1}(0), g|_{f^{-1}(0)}\right)$. Thus the following result is obtained.

Theorem 14.47 (Garcia–Rio and Kupeli). *Let (M, g) be a stably causal space–time with smooth time function $f : M \to \mathbb{R}$. Let*

$$g_c = -g(\nabla(f), \nabla(f)) \cdot g.$$

Then either

(1) *(M, g_c) is timelike geodesically incomplete, or*
(2) *(M, g) is conformally diffeomorphic to a type of parametrized warped product $(\mathbb{R} \times N, -dt^2 \oplus g(t))$, where $N = f^{-1}(0)$ and $g(t)$, $t \in \mathbb{R}$, is a one-parameter family of Riemannian metrics for N.*

APPENDIX A

JACOBI FIELDS AND TOPONOGOV'S THEOREM FOR LORENTZIAN MANIFOLDS†

One of the important consequences of the Rauch Comparison Theorem in Riemannian geometry is the Toponogov Triangle Comparison Theorem [cf. Cheeger and Ebin (1975, pp. 42–49)]: Let M be a complete Riemannian manifold (metric here and elsewhere in this appendix denoted by $\langle \cdot, \cdot \rangle$) with sectional curvature of all 2-planes σ in M satisfying $K(\sigma) \geq H$ for some number H. Let $(\gamma_1, \gamma_2, \gamma_3)$ be geodesics in M forming a triangle: $\gamma_1(0) = \gamma_2(0)$, $\gamma_1(L_1) = \gamma_3(0)$, and $\gamma_2(L_2) = \gamma_3(L_3)$, where $L_i = L(\gamma_i)$. Suppose that γ_2 and γ_3 are minimal geodesics and, in the case that $H = q^2$ $(q > 0)$, suppose that $L_i \leq \pi/q$ for $i = 1, 2, 3$. Assume the geodesics γ_i are parametrized by arc length, and define $\alpha_3 = \langle \gamma_1'(0), \gamma_2'(0) \rangle$ and $\alpha_2 = \langle -\gamma_1'(L_1), \gamma_3'(0) \rangle$. For a triangle $(\overline{\gamma}_1, \overline{\gamma}_2, \overline{\gamma}_3)$, possibly in another manifold, $\overline{\alpha}_2$ and $\overline{\alpha}_3$ are defined similarly.

(1) In the simply connected two-dimensional Riemannian manifold M_H of constant curvature H, there is a geodesic triangle $(\overline{\gamma}_1, \overline{\gamma}_2, \overline{\gamma}_3)$ with $L(\overline{\gamma}_i) = L_i$, $i = 1, 2, 3$, and $\overline{\alpha}_2 \leq \alpha_2$, $\overline{\alpha}_3 \leq \alpha_3$. This triangle is determined up to congruence if $H \leq 0$, or if $H > 0$ and all $L_i < \pi/q$.

(2) In M_H, let $\overline{\gamma}_1$ and $\overline{\gamma}_2$ be geodesics with $\overline{\gamma}_1(0) = \overline{\gamma}_2(0)$, $L(\overline{\gamma}_1) = L_1$, $L(\overline{\gamma}_2) = L_2$, and $\overline{\alpha}_3 = \alpha_3$. Let $\overline{\gamma}_3$ be a minimal geodesic between the endpoints of $\overline{\gamma}_1$ and $\overline{\gamma}_2$. Then $L(\overline{\gamma}_3) \geq L(\gamma_3)$.

It is possible to develop an analogous theorem for Lorentzian geometry; the program will be sketched here, although the proofs will be omitted. Details appear in Harris (1979, pp. 3–41) and Harris (1982a).

†By Steven G. Harris, Department of Mathematics, St. Louis University, St. Louis, Missouri

The first step is to modify the Timelike Rauch Comparison Theorem I (cf. Theorem 11.11) so that it applies to timelike geodesics, not without conjugate points, but without focal points: If N is a submanifold of a manifold M, then a point $q \in M$ will be said to be a *focal point of* N *from* p ($p \in N$) if q is the image of a critical point of exp in the normal bundle of N at p. For v a vector in the tangent bundle TM, $N(v)$ will denote the submanifold of M which is the image under exp of a small enough neighborhood of the origin in the perpendicular space of v such that exp is an embedding on it. Thus $N(v)$ is an $(n-1)$-dimensional submanifold orthogonal to v.

In the following statement of the second Rauch Theorem, and elsewhere in this appendix, $A \wedge B$ will denote the 2-plane spanned by the vectors A and B.

Theorem A.1 (Timelike Rauch II). *Let V_i be a Jacobi field along a unit speed timelike geodesic γ_i in a space–time M_i, $i = 1, 2$, with $\gamma_i : [0, L] \to M_i$; let $T_i = \gamma_i'$. Suppose that $\langle V_1, V_1\rangle_0 = \langle V_2, V_2\rangle_0$, $\langle V_1, T_1\rangle_0 = \langle V_2, T_2\rangle_0$, and $(\nabla_{T_i} V_i)(0) = 0$. Further suppose that for any vectors X_i at $\gamma_i(t)$*

$$K(X_1 \wedge T_1) \geq K(X_2 \wedge T_2)$$

and that γ_2 has no focal point of $N(\gamma_2'(0))$ from $\gamma_2(0)$. Then for all t in $[0, L]$,

$$\langle V_1, V_1\rangle_t \geq \langle V_2, V_2\rangle_t.$$

An important corollary of this theorem shows how curvature can affect the lengths of timelike curves.

Corollary A.2. *Let $\gamma_i : [0, L] \to M_i$, $i = 1, 2$, be two timelike (or two null or two spacelike) geodesics, and let E_i be parallel unit timelike vector fields along γ_i with $\langle E_1, T_1\rangle = \langle E_2, T_2\rangle$ ($T_i = \gamma_i'$). Let $f : [0, L] \to \mathbb{R}$ be any smooth real-valued function. Suppose that for all t in $[0, L]$, $\exp_{\gamma_i(t)}(f(t)E_i(t))$ is defined; call this $c_i(t)$. Suppose further that for all t in $[0, L]$, the geodesic $\eta : s \mapsto \exp_{\gamma_2(t)}(sE_2(t))$ has no focal point of $N(\eta'(0))$ from $\eta(0)$ for $s \leq f(t)$. Finally, suppose that for all timelike 2-planes σ_i in M_i,*

$$K(\sigma_1) \geq K(\sigma_2).$$

Then, for all t in $[0, L]$,

$$\langle {c_1}', {c_1}' \rangle_t \geq \langle {c_2}', {c_2}' \rangle_t.$$

Thus if c_1 is a nonspacelike curve, then c_2 is also nonspacelike, and

$$L(c_1) \leq L(c_2).$$

This corollary makes possible a triangle comparison theorem for "thin" triangles. The model spaces used for comparison are the two-dimensional de Sitter and anti–de Sitter spaces of constant curvature [cf. Hawking and Ellis (1973, pp. 124–134), Wolf (1961, pp. 114–118)]. A triangle of timelike geodesics $(\gamma_1, \gamma_2, \gamma_3)$ is "thin" in this context if the following holds: For a given H, let $\overline{\gamma}_1$ and $\overline{\gamma}_2$ be timelike geodesics in the simply connected two-dimensional Lorentzian manifold of constant curvature H (denoted by M_H) with $\overline{\gamma}_1(0) = \overline{\gamma}_2(0)$, $L(\overline{\gamma}_1) = L_1$, $L(\overline{\gamma}_2) = L_2$, and $\overline{\alpha}_3 = \alpha_3$. First, suppose there is a timelike geodesic $\overline{\gamma}_3$ between the endpoints of $\overline{\gamma}_1$ and $\overline{\gamma}_2$. Let $\overline{E}$ be the parallel translate of ${\overline{\gamma}_1}'(0)$ along $\overline{\gamma}_2$. For each t in $[0, L_2]$, there is a smallest positive number $f(t)$ such that $\exp(f(t)\overline{E}(t))$ lies on $\overline{\gamma}_3$. Second, suppose that for all such t, the geodesic $s \mapsto \exp(s\overline{E}(t))$ has no focal point of $N(\overline{E}(t))$ from $\overline{\gamma}_2(t)$ for $s \leq f(t)$. If these two suppositions hold, and if γ_3 is maximal and all timelike planes σ in M satisfy $K(\sigma) \leq H$, then $L(\gamma_3) \geq L(\overline{\gamma}_3)$.

The problem is to start with a more general timelike geodesic triangle, slice it up into "thin" triangles, apply the result just mentioned to each slice, and then put them back together. In the Riemannian theorem, completeness is used to ensure that in any triangle $(\gamma_1, \gamma_2, \gamma_3)$, minimal geodesics can be extended from $\gamma_3(L_3)$ to γ_1, slicing up the original triangle. In the Lorentzian context, global hyperbolicity succeeds just as well as long as $H \geq 0$. For $H = -q^2$, however, a problem arises: Not even the model spaces M_H are globally hyperbolic; indeed, by Proposition 11.8, no timelike geodesic of length greater than π/q can be maximal in a Lorentzian manifold whose timelike planes σ satisfy $K(\sigma) \leq -q^2$. A solution to this problem lies in a new concept, a sort of global hyperbolicity in the small: For x and y in a Lorentzian manifold M, let $C(x, y)$ denote the space of nonspacelike curves in M from x to y, modulo reparametrization, with the compact-open topology.

Definition A.3. A Lorentzian manifold M is *globally hyperbolic of order* q ($q > 0$) if M is strongly causal and for all points x and y in M with $\sup\{L(\gamma) : \gamma \in C(x,y)\} < \pi/q$, this space $C(x,y)$ is compact.

The Lorentzian analogue of Toponogov's Theorem can now be stated (all geodesics parametrized with unit speed).

Theorem A.4 (Lorentzian Triangle Comparison Theorem). *Let M be a space–time whose timelike planes σ satisfy $K(\sigma) \leq H$ for some constant H; M is to be globally hyperbolic or, in case $H = -q^2$, globally hyperbolic of order q. Let $(\gamma_1, \gamma_2, \gamma_3)$ be a triangle of timelike geodesics with γ_2 the future directed side between the pastmost and futuremost of the three endpoints, γ_1 the other future directed side from $\gamma_2(0)$, and γ_3 the remaining future directed side. Suppose that γ_2 and γ_3 are maximal and, if $H = -q^2$, that $L_i < \pi/q$, $i = 1, 2, 3$. Then*

(1) *There is in M_H a timelike geodesic triangle $(\overline{\gamma}_1, \overline{\gamma}_2, \overline{\gamma}_3)$ with $L(\overline{\gamma}_i) = L_i$, $i = 1, 2, 3$, and with $\overline{\alpha}_2 \geq \alpha_2$ and $\overline{\alpha}_3 \geq \alpha_3$.*

(2) *For timelike geodesics $\overline{\gamma}_1$ and $\overline{\gamma}_2$ constructed in M_H with $\overline{\gamma}_1(0) = \overline{\gamma}_2(0)$, $L(\overline{\gamma}_1) = L_1$, $L(\overline{\gamma}_2) = L_2$, and $\overline{\alpha}_3 = \alpha_3$, if there is a timelike geodesic $\overline{\gamma}_3$ between the endpoints of $\overline{\gamma}_1$ and $\overline{\gamma}_2$, then $L(\overline{\gamma}_3) \leq L(\gamma_3)$.*

The Toponogov Triangle Comparison Theorem can be used to show how a bound on sectional curvature can rigidly determine a complete Riemannian manifold if the limits of the curvature-imposed constraints are attained. One such result is the maximal diameter theorem [cf. Cheeger and Ebin (1975, p. 110)]: If a complete Riemannian manifold M^n satisfies $K(\sigma) \geq q^2 > 0$ for all planes σ, and if the diameter of M attains the maximum thus allowable, π/q, then M is isometric to the sphere S^n of curvature q^2.

Theorem A.4 can be used similarly.

Theorem A.5. *Let M^n be a space–time which is globally hyperbolic of order q and satisfies $K(\sigma) \leq -q^2$ for all timelike planes σ. Suppose that M possesses a complete timelike geodesic γ which is maximal on all intervals of length π/q. Then M is isometric to the simply connected geodesically complete n-dimensional Lorentzian manifold of constant curvature $-q^2$.*

In addition to timelike geodesics, it is possible to study the effects of curvature on Jacobi fields along null geodesics. Sectional curvature cannot be used for this since it is undefined for null (singular) planes, but a similar concept, introduced here, will work.

Definition A.6. If σ is a null plane and N is any nonzero element of the one-dimensional space of null vectors in σ, then the *null sectional curvature of σ with respect to N*, $K_N(\sigma)$, is defined by

$$K_N(\sigma) = \frac{\langle R(A,N)N, A\rangle}{\langle A, A\rangle}$$

where A is any nonnull vector in σ.

This expression for null sectional curvature is independent of the vector A and depends in a quadratic fashion on the vector N. This curvature quantity has certain interesting relations with ordinary sectional curvature. For instance, see Proposition 2.3 in Harris (1982a).

Proposition A.7. *If at a single point in a Lorentzian manifold the null sectional curvatures are all positive (respectively, negative), then timelike sectional curvature at that point is unbounded below (respectively, above); and if the null sectional curvatures all vanish, then the timelike and spacelike sectional curvatures are all equal. Thus, a Lorentzian manifold of dimension at least three has constant curvature iff it has null sectional curvature everywhere zero.*

Null sectional curvature can be used much like timelike sectional curvature with regard to Jacobi fields.

Proposition A.8. *Let $\beta : [0, L] \to M^n$ be a null (affinely parametrized) geodesic with $T = \beta'$. If for all nonnull vectors V perpendicular to T, $K_T(V \wedge T) \leq 0$, then β has no conjugate points. If for some q, $K_T(V \wedge T) \geq q^2$ for all such V or, more generally, if $\mathrm{Ric}(T,T) \geq (n-2)q^2$, then $L \geq \pi/q$ implies β does have a conjugate point.*

Null sectional curvature also lends itself to a Rauch-type theorem.

Theorem A.9 (Rauch Comparison Theorem for Null Geodesics). *Let* $\beta_i : [0, L] \to M_i$, $i = 1, 2$, *be null geodesics;* $T_i = \beta_i{}'$. *Let* V_i, $i = 1, 2$, *be perpendicular Jacobi fields along* β_i, *not everywhere parallel to* T_i, *(and thus nowhere parallel to* T_i*), with* $V_i(0) = 0$ *and* $\langle V_1{}', V_1{}' \rangle_0 = \langle V_2{}', V_2{}' \rangle_0$.

Suppose that for any t *in* $[0, L]$ *and for any nonnull vectors* X_i *at* $\beta_i(t)$, *perpendicular to* T_i,

$$K_{T_1}(X_1 \wedge T_1) \le K_{T_2}(X_2 \wedge T_2)$$

and that β_2 *has no points conjugate to* $\beta_2(0)$. *Then for all* t *in* $[0, L]$,

$$\langle V_1, V_1 \rangle_t \ge \langle V_2, V_2 \rangle_t.$$

The necessity of using perpendicular vector fields makes this theory less tractable than that for timelike geodesics.

Details for Proposition A.8 and Theorem A.9 can be found in Harris (1982a).

Finally, in Harris (1985), consideration is given to the case where the null curvature function of all null planes associated to a null congruence is a point function. It is shown that this condition locally characterizes Robertson–Walker metrics and that if further completeness and causality assumptions are added, then a global characterization may be obtained.

APPENDIX B

FROM THE JACOBI, TO A RICCATI, TO THE RAYCHAUDHURI EQUATION: JACOBI TENSOR FIELDS AND THE EXPONENTIAL MAP REVISITED

The use of Jacobi field techniques in the so-called "comparison theory" in Riemannian geometry stems from the close relationship between Jacobi fields and the differential of the exponential map. One very basic example is provided by Proposition 10.16 in Section 10.1, in which the Jacobi field J along a unit timelike geodesic c with $J(0) = 0$ and $J'(0) = w$ is given by

$$J(t) = \exp_{p*}\left(t\,\tau_{tc'(0)}w\right).$$

In this appendix we give a second illustration of the relationship between the differential of the exponential map and Jacobi fields which has come to prominence in the 1980's version of comparison theory in Riemannian geometry. In this comparison theory, the use of the Jacobi equation has been partly supplanted by the use of a Riccati equation for the second fundamental tensor of a local foliation of a Riemannian manifold by hypersurfaces arising as the level sets of a differentiable function with gradient of constant length [cf. Karcher (1989), Eschenburg (1975, 1987, 1989)]. A recent application of the Riccati inequality method to submanifolds of semi-Riemannian manifolds is given in Andersson and Howard (1994).

B.1 Generalities on Semi-Riemannian Manifolds

Let (M, g) be a semi-Riemannian manifold with semi-Riemannian metric g of arbitrary signature (s, r) but which is nondegenerate. Let U be an open subset of M. This section is concerned with the basic differential geometry of the foliation induced on U by the level sets of a smooth function $f : U \to \mathbb{R}$ with nowhere vanishing gradient vector field $\nabla f = \operatorname{grad} f$ of everywhere constant length. Upon rescaling, it may be supposed that

$$g(\operatorname{grad} f, \operatorname{grad} f) = \lambda \tag{B.1}$$

for some choice of λ in $\{-1, 0, +1\}$. Such functions have been termed solutions to the eikonal equation.

This formalism has been useful in a number of different geometric contexts. The first and perhaps most well-known example occurs in the case where (M, g) is Riemannian, U is a deleted neighborhood of p which does not intersect the cut locus of p, and $f = d(p, \cdot)$, where d denotes the Riemannian distance function of (M, g). In this case, $\lambda = 1$ in (B.1). The analogous situation may be considered on a strongly causal space–time where U could be taken to be a subset of $I^+(p)$ which does not meet the future timelike cut locus of p, and $f = d(p, \cdot)$, where d this time denotes the Lorentzian distance function of (M, g). In this case, $\lambda = -1$. The case $\lambda = -1$ also arises (a posteriori) in the course of the proof of the Lorentzian splitting theorem, as in Beem, Ehrlich, Markvorsen, and Galloway (1985, p. 39), in which U could be taken to be a tubular neighborhood of a complete timelike geodesic line $c : \mathbb{R} \to (M, g)$ in the globally hyperbolic space–time (M, g), and $f = b^+$ (or $f = b^-$) is the Busemann function associated to the future timelike ray (respectively, the past timelike ray) determined by the given timelike line. The third case, $\lambda = 0$, arises in the context of semi-time functions for space–times, i.e., $g(\operatorname{grad} f, \operatorname{grad} f) \leq 0$ is required for a semi-time function. The case that $\operatorname{grad} f$ is identically null is encountered for the space–time $M = \mathbb{R} \times S^1$ with Lorentzian metric $ds^2 = d\theta\, dt$ and $f(t, \theta) = -t$. A second larger class of examples is furnished by the gravitational plane wave space–times. Here the global coordinate function $f(x, y, u, v) = u$ has gradient $\operatorname{grad} f = \partial/\partial v$, which is a global null parallel field. The geometric properties of this so-called "quasi-time function" have been systematically studied for these space–times in Ehrlich and Emch (1992a,b, 1993) [cf. Chapter 13].

The following basic result explains the significance of constant gradient length in (B.1).

Lemma B.1. *Let* $\lambda = -1$ *(respectively,* $+1$, 0*). Suppose* $f : U \to \mathbb{R}$ *is a smooth function with* $g(\nabla f, \nabla f) = \lambda$*. Let* $c : [a, b] \to U$ *be an integral curve of* $\operatorname{grad} f$*. Then* c *is a timelike (respectively, spacelike, null) geodesic.*

Proof. Fix any p in U and v in T_pM. Denoting the Hessian $D(Df)$ by H^f, we have

$$\begin{aligned} g\left(D_{\nabla f}\nabla f, v\right) &= H^f(\nabla f, v) = H^f(v, \nabla f) \\ &= g(D_v\nabla f, \nabla f) = \tfrac{1}{2}v(g(\nabla f, \nabla f)) \\ &= \tfrac{1}{2}v(\lambda) = 0. \end{aligned}$$

Since v was arbitrary and the semi-Riemannian metric was assumed to be nondegenerate, we have

$$D_{\nabla f}\nabla f = 0 \tag{B.2}$$

on U. Thus the integral curves of $\operatorname{grad} f$ are geodesics. □

In the case that (M, g) is Lorentzian and $\lambda = -1$, or in the case that (M, g) is Riemannian and $\lambda = +1$, it is standard that (B.1) has the even stronger implication that any such integral curve of $\operatorname{grad} f$ is a maximal geodesic in the space–time $(U, g|_U)$ (respectively, a minimal geodesic in the Riemannian manifold $(U, g|_U)$) by virtue of the reverse (or wrong way) Cauchy–Schwarz inequality for timelike vectors [cf. Sachs and Wu, (1977a)] (respectively, the Cauchy–Schwarz inequality for tangent vectors).

We present here the argument for the space–time case, under the assumption that the vector field $\operatorname{grad} f$ is future directed. Thus let $\gamma : [a, b] \to U$ be any future directed nonspacelike curve in U with $\gamma(a) = c(t_1)$ and $\gamma(b) = c(t_2)$, where c is a maximal integral curve of $\operatorname{grad} f$ lying in U. Then we first have that

$$g(\operatorname{grad} f, \gamma') = \frac{d(f \circ \gamma)}{dt} < 0$$

since both vectors are future directed and $\operatorname{grad} f$ is timelike. Secondly, by the wrong-way Cauchy–Schwarz inequality in the case that $\gamma'(t)$ is timelike, and trivially if $\gamma'(t)$ is null, we have

$$\begin{aligned} |g(\gamma', \operatorname{grad} f)| &\geq |g(\gamma', \gamma')|^{\frac{1}{2}}\, |g(\operatorname{grad} f, \operatorname{grad} f)|^{\frac{1}{2}} \\ &= |g(\gamma', \gamma')|^{\frac{1}{2}}\,, \end{aligned}$$

so that

$$
\begin{aligned}
L(\gamma) &= \int_{t=a}^{b} |-g(\gamma', \gamma')|^{\frac{1}{2}} \, dt \\
&\le \int_{a}^{b} |g(\gamma', \operatorname{grad} f)| \, dt \\
&= \int_{a}^{b} \frac{-d(f \circ \gamma)}{dt} \, dt = f(\gamma(a)) - f(\gamma(b)) \\
&= f(c(t_1)) - f(c(t_2)) = |t_2 - t_1|
\end{aligned}
$$

as required. The argument for the Lorentzian case with grad f null is deferred until after Lemma B.3.

Let us now record an elementary identity in the context of a differentiable map $f : U \to \mathbb{R}$ for U an open subset of the (nondegenerate) semi-Riemannian manifold (M, g) that will be useful in the sequel:

$$
\text{(B.3)} \qquad f_{*p}(v) = g(v, \operatorname{grad} f|_p) \left. \frac{\partial}{\partial t} \right|_{f(p)} \qquad \text{in } T_{f(p)}\mathbb{R}
$$

for any p in U and v in T_pM. This identity follows immediately from the representation

$$
v = \sum_i v(x_i) \frac{\partial}{\partial x_i}
$$

applied to the chart $t(r) = r$ for $\mathbb{R}$ and the tangent vector $f_{*p}(v)$, i.e.,

$$
\begin{aligned}
f_{*p}v &= (f_{*p}v)(\mathrm{Id}) \left. \frac{\partial}{\partial t} \right|_{f(p)} = v(\mathrm{Id} \circ f) \left. \frac{\partial}{\partial t} \right|_{f(p)} \\
&= v(f) \left. \frac{\partial}{\partial t} \right|_{f(p)} = g(\nabla f(p), v) \left. \frac{\partial}{\partial t} \right|_{f(p)} .
\end{aligned}
$$

The identity (B.3) then has the following consequence.

Lemma B.2. *For λ in $\{-1, 0, +1\}$, let $f : U \to \mathbb{R}$ have constant gradient length λ as in (B.1). Then $\mathrm{Im}(f_{*p}) = T_{f(p)}\mathbb{R}$ for each p in U. Hence, f is a submersion and $\dim(f^{-1}(r)) = \dim(M) - 1$ for each r in $f(U)$ in all three cases simultaneously.*

Proof. This is an immediate consequence of the assumed nondegeneracy of the given semi-Riemannian metric and (B.3). For the only way that f_{*p} can fail to be surjective is that $\mathrm{Im}(f_{*p}) = \{0\}$, since $\dim(T_{f(p)}\mathbb{R}) = 1$. But in this case, we then have from (B.3) that $g(v, \operatorname{grad} f(p)) = 0$ for all v in T_pM. But this then implies that $\operatorname{grad} f(p) = 0$, in contradiction to the assumption that $\operatorname{grad} f$ is nowhere vanishing. □

The following is the translation to the semi-Riemannian context of a basic result in elementary differential topology.

Lemma B.3. *Let $c : [0, a) \to U$ be an integral curve of* $\operatorname{grad} f$ *with $c(0)$ in $f^{-1}(r)$. Then*

(1) *If $\lambda = 0$, then $c(t) \in f^{-1}(r)$ for all t;*
(2) *If $\lambda = +1$, then $c(t) \in f^{-1}(r + t)$ for all t; and*
(3) *If $\lambda = -1$, then $c(t) \in f^{-1}(r - t)$ for all t.*

Proof. Since $c(t)$ is an integral curve of $\operatorname{grad} f$, we have from identity (B.3) that

$$\begin{aligned}(f \circ c)'(t) &= f_* c'(t) = g\big(c'(t), \operatorname{grad} f(c(t))\big) \\ &= g(\nabla f(c(t)), \nabla f(c(t))) = \lambda.\end{aligned}$$

Hence if $\lambda = 0$, then $(f \circ c)'(t) = 0$ for all t and (1) follows. The other cases are similar. (If $\lambda = -1$, then $(f \circ c)'(t) = -1$ so that $[f \circ c](t) = r - t$.) □

Thus in the case of a null gradient, instead of mapping a level surface to a different level surface, the gradient flow of f maps each level surface into itself. This phenomenon is encountered for the gravitational plane wave space–times where $f(y, z, u, v) = u$ is the "natural" quasi-time function for this class of space–times and has everywhere nonvanishing but null gradient (cf. Chapter 13).

Also as mentioned above, we may derive the following consequence from Lemma B.3. Let (M, g) be Lorentzian, and let $c : (a, b) \to U$ be an integral curve of $f : U \to \mathbb{R}$ with $\operatorname{grad} f = 0$. Then c is globally maximal with respect to arc length among future directed nonspacelike curves lying in U. For by considering $-\operatorname{grad} f$ if necessary, we may assume that c is future directed. If c fails to be maximal in the above sense, then there exist s, t with $a < s < t < b$

and a future directed timelike curve $\sigma : [0,1] \to U$ with $\sigma(0) = c(s)$ and $\sigma(1) = c(t)$ (cf. Section 9.2). But then $(f \circ \sigma)'(u) = \langle \nabla f(\sigma(u)), \sigma'(u) \rangle > 0$, so that $f(c(s)) = f(\sigma(0)) > f(\sigma(1)) = f(c(t))$, in contradiction to part (1) of Lemma B.3.

Related to Lemmas B.2 and B.3, and also a consequence of equation (B.3), is the following description of the tangent spaces and normal vector field to the local foliation $\{f^{-1}(u) : u \in \mathbb{R}\}$ induced on U by the level submanifolds of f.

Lemma B.4. *For any λ in $\{-1, 0, +1\}$, we have simultaneously*

(1) *The tangent space at any p in $f^{-1}(r)$ is given by*

$$T_p(f^{-1}(r)) = \ker(f_{*p}) = \{v \in T_pM : g(v, \operatorname{grad} f(p)) = 0\}; \tag{B.4}$$

(2) *$N = \operatorname{grad} f$ is a normal field of constant length for each level submanifold $f^{-1}(r)$; and*

(3) *For any v in $T_p(f^{-1}(r))$ we have $D_v \operatorname{grad} f \in T_p(f^{-1}(r))$.*

Proof. (3) In view of (B.3) it suffices to note that (as before)

$$g(D_v \nabla f, \nabla f) = v(g(\nabla f, \nabla f)/2) = v(\lambda)/2 = 0. \quad \square$$

In the case that (M, g) is Lorentzian and $\operatorname{grad} f$ is timelike (respectively, spacelike), the induced metric on the level submanifolds is nondegenerate and the submanifold is thus spacelike (respectively, timelike) in the sense of Definition 3.47. In the case that $\operatorname{grad} f$ is null, the induced metric on the level submanifold is degenerate in view of (B.4) since $\operatorname{grad} f(p)$ is both in $T_p(f^{-1}(f(p)))$ and orthogonal to all vectors in $T_p(f^{-1}(f(p)))$. Hence, in the case that $\operatorname{grad} f$ is null, the level submanifolds are *not* "semi-Riemannian submanifolds" in the sense of O'Neill (1983), so that some care needs to be taken in considering the concepts of the second fundamental form and of a totally geodesic submanifold.

To be consistent with Section 3.5 in notation, given $\operatorname{grad} f$ on U we define an operator L on $T_p(f^{-1}(r))$ by

$$L(v) = -D_v \operatorname{grad} f \qquad \text{in } T_pM. \tag{B.5}$$

In the case that $\lambda = +1$ or $\lambda = -1$ and (M, g) is Lorentzian, since the level submanifolds inherit nondegenerate induced metrics, it is well known that

$$L : T_p(f^{-1}(r)) \to T_p(f^{-1}(r)) \tag{B.6}$$

is the shape operator whose vanishing determines whether the level submanifold is totally geodesic. In the case that $\lambda = 0$ (or (M, g) is not Lorentzian), it is a consequence of part (3) of Lemma B.4 rather than familiar abstract nonsense that (B.6) is still valid, so we may continue to regard L as the shape operator. This is further confirmed by the following lemma which is standard when the induced metric is nondegenerate.

Lemma B.5. *Let (M, g) be an arbitrary (nondegenerate) semi-Riemannian manifold, and let $\lambda \in \{-1, 0, +1\}$. Suppose $f : U \to \mathbb{R}$ satisfies the further condition*

$$D_v \operatorname{grad} f = 0 \tag{B.7}$$

for all v in T_pM and for all p in U. Then every level submanifold $f^{-1}(r)$ in U is totally geodesic in the following sense: Let $v \in T_p(f^{-1}(r))$, and let $c : (a, b) \to U$ be (the restriction of) the geodesic in (M, g) with $c'(0) = v$. Then $c((a, b))$ is contained in $f^{-1}(r)$.

Proof. Consider the function $F : (a, b) \to \mathbb{R}$ given by $F(t) = f \circ c(t)$. Then we have $F(0) = r$, and we need to show that $F'(t) = 0$ for all t to obtain the desired result. Now $F'(t) = (f \circ c)'(t) = g(\nabla f(c(t)), c'(t))$ and thus $F'(0) = g(\nabla f(p), v) = 0$ by (B.4). Further,

$$\begin{aligned} F''(t) &= \frac{d}{dt}\big(g(\nabla f(c(t)), c'(t))\big) \\ &= g(D_{c'(t)} \nabla f, c'(t)) + g(\nabla f(c(t)), D_{c'(t)} c'(t)) = 0 \end{aligned}$$

since c is a geodesic and (B.7) is assumed. Thus $F'(t) = F'(0) = 0$ as required. $\square$

B.2 Jacobi Tensors and the Normal Exponential Map in the Timelike Case

In this section, we will assume that (M, g) is Lorentzian and that U is equipped with a smooth function $f : U \to \mathbb{R}$ with everywhere unit timelike gradient, i.e., $g(\operatorname{grad} f, \operatorname{grad} f) = -1$. Since $\operatorname{grad} f$ is normal to the level surfaces of f, we will denote $N = \operatorname{grad} f$. Also, since we are only interested in local calculations in this section, we will not bother to carefully specify domains of geodesics or of maps in the t or s-parameters.

Fix r in $\mathbb{R}$ with r in $f(U)$. With the above proviso in mind, recall that it is traditional to study the geometry of the submanifold $f^{-1}(r)$ by considering the so-called normal exponential map $E(\,\cdot\,, t)$ on $f^{-1}(r)$, which may be defined by

$$E(p, t) = \exp_p(tN(p)). \tag{B.8}$$

Since the integral curves of $N = \operatorname{grad} f$ are geodesics, it follows that in this particular geometric situation (B.8) is equivalent to following the gradient flow of f starting at points of $f^{-1}(r)$, and thus Lemma B.3 implies that $E(\,\cdot\,, t)$ is a map

$$E(\,\cdot\,, t) : f^{-1}(r) \to f^{-1}(r - t). \tag{B.9}$$

The normal exponential map may be placed in a variational context with notation as in Chapter 10 for the purpose of studying Jacobi fields along timelike geodesics initially perpendicular to $f^{-1}(r)$ as in Karcher (1989) for Riemannian manifolds.

Fix p in $f^{-1}(r)$, and let $c : [0, a) \to M$ denote the integral curve of N with $c(0) = p$; by the above comments, $c(t) = \exp(tN(p))$. Now let $v \in T_p(f^{-1}(r))$ be given, and choose any smooth curve $b : (-\epsilon, \epsilon) \to f^{-1}(r)$ with $b(0) = p$ and $b'(0) = v$ for purposes of computation. Then define a variation $\alpha(t, s)$ of c by

$$\alpha(t, s) = E(b(s), t) = \exp_{b(s)}(tN(b(s))) \tag{B.10}$$

so that

$$\alpha(t, \,\cdot\,) : f^{-1}(r) \to f^{-1}(r - t)$$

and

$$\alpha_* \, \partial/\partial t = \operatorname{grad} f \circ \alpha.$$

Since $s \mapsto \alpha(\,\cdot\,, s)$ is a variation of c through geodesics, the variation vector field

$$J(t) = \alpha_* \left.\frac{\partial}{\partial s}\right|_{(t,0)} = \left.\frac{d(E(b(s),t))}{ds}\right|_{(t,0)} \tag{B.11}$$

is a Jacobi field. By using the chain rule or by introducing some fancy notation, one may rewrite the third term of equation (B.11) in a form closer to that of Proposition 10.16 as recalled in the introduction to this Appendix. Explicitly, if we let

$$i : \mathbb{R} \times \mathbb{R} \to f^{-1}(r) \times \mathbb{R}$$

be the map given by $i(s,t) = (b(s),t)$, then $\alpha = E \circ i$ and $i_* \left.\right|_{(t,0)} \frac{\partial}{\partial s} = (b'(0), 0) = (v, 0)$. Hence, we have

$$J(t) = E_*|_{(p,t)} \, (v, 0) \tag{B.12}$$

for all t in $[0, a)$. Hence, the set of Jacobi fields $J(t)$ as v varies over all vectors tangent to $f^{-1}(r)$ at p is identified with the differential E_* of the normal exponential map E by means of equation (B.12).

Lemma B.6. *The variation vector field $J(t)$ of the variation $\alpha(t,s)$ given by (B.10) satisfies the differential equation*

$$J'(t) = -L_t(J(t)) \tag{B.13}$$

for all t in $[0, a)$ with initial conditions $J(0) = v$ and $J'(0) = -L(v)$, where L is the second fundamental tensor for the level surfaces $\{f^{-1}(r - t)\}$ defined as in (B.5) by $L_t(w) = -D_w N$.

Proof. Using (B.11), one calculates

$$\begin{aligned} J'(t) &= D_{\partial/\partial t} \left(\alpha_* \tfrac{\partial}{\partial s}\right)\big|_{(t,0)} \\ &= D_{\partial/\partial s} \left(\alpha_* \tfrac{\partial}{\partial t}\right)\big|_{(t,0)} \\ &= D_{\partial/\partial s} \operatorname{grad} f \circ \alpha \\ &= D_{\alpha_*(\partial/\partial s)(t,0)} \operatorname{grad} f \\ &= D_{J(t)} N = -L_t(J(t)). \quad \square \end{aligned}$$

Thus in our particular geometric situation, the usual submanifold initial condition for Jacobi fields, $J'(0) = -L(v) = -L(J(0))$, is propagated for all t along the level submanifolds $f^{-1}(r-t)$ and the transversal timelike geodesic c. Further, we may reinterpret equation (B.13) in the language of Jacobi tensor fields as utilized in Section 10.3 and Sections 12.2–12.3. Let

$$A = A(t) : N(c(t)) \to N(c(t)) = (c'(t))^{\perp} \tag{B.14}$$

be the Jacobi tensor along $c : [0, a) \to M$ satisfying the initial conditions

$$A(0) = E \qquad \text{and} \qquad A'(0) = -L. \tag{B.15}$$

Then as we remarked in Lemma 12.21, A is a Lagrange tensor field which intuitively represents the vector space of Jacobi fields J along c which satisfy the initial condition $J'(0) = -L(J(0))$. In view of equation (B.13), we have in the present context the further identification for the associated tensor field $B = A'A^{-1}$ occurring in the Raychaudhuri equation that

$$B(t) = -L(t) \tag{B.16}$$

for all t in $[0, a)$ for which $A(t)$ is non-singular, i.e., which are not focal points to p. Also, one may interpret equation (B.12) as identifying the differential E_* of the normal exponential map with the Jacobi tensor A. Equation (B.16) then has the implication that the Raychaudhuri equation for A (or equivalently, E_*) is related to the trace of L and its derivative L', hence to the mean curvature of the level surfaces $\{f^{-1}(r-t)\}$ (cf. equation (12.2) in Section 12.1).

B.3 The Riccati Equation for L

There is a venerable technique in the disconjugacy theory of ordinary differential equations in which a Jacobi equation such as $u'' + q(t)u = 0$ is linked to an associated Riccati equation $r' + r^2 + q(t) = 0$ by the change of variables $r = u'/u$ [cf. Hartman (1964)]. The $n \times n$ matrix version of this theory involves forming $B = A'A^{-1}$ exactly as is done in general relativity in passing from the Jacobi equation $A'' + RA = 0$ to the associated Raychaudhuri equation (cf. Sections 10.3 and 12.2–12.3).

In this section, we shall follow the route suggested in Karcher (1989) and obtain an intermediate Riccati equation satisfied by the second fundamental form operator in the geometric context under consideration. Thus, let (M, g) be a Lorentzian manifold and $f : U \to \mathbb{R}$ be a smooth function with everywhere timelike gradient of constant length -1, as in Section B.2. Let $p \in f^{-1}(r)$ be fixed, and let $c : [0, a) \to U$ denote the integral curve of f starting at p. Choose any $v \in T_p(f^{-1}(r))$, and let $J = J_v$ denote the Jacobi field along c with initial conditions $J(0) = v$ and $J'(0) = -L(v)$. We established in Section B.2 the identity

$$J'(t) = -L \circ J\,(t) \tag{B.17}$$

for all t in $[0, a)$. Recalling here that $X'(t) = D_{\partial/\partial t}X$ for vector fields and tensors along the geodesic c, we have

$$\left[D_{\partial/\partial t}L\right](J(t)) = D_{\partial/\partial t}(L(J)) - L\left(D_{\partial/\partial t}J\right),$$

or in more compact notation,

$$L'(J) = (L \circ J)' - L(J').$$

Thus differentiating (B.17) covariantly with respect to t yields

$$\begin{aligned} J'' &= -(L \circ J)' = -L'(J) - L(J') \\ &= -L'(J) - L(-L(J)) \end{aligned}$$

again using (B.17). Combining this last result with the Jacobi equation $J'' + R(J, c')c' = 0$ then yields the desired Riccati type equation for J_v:

$$L'(J_v) - (L \circ L)(J_v) = R(J_v, c')c' \qquad \text{for all } t \text{ in } [0, a). \tag{B.18}$$

If we denote the composition $L \circ L$ by L^2, then prior to the first focal point to $t = 0$ along c to the submanifold $f^{-1}(r)$, we have from (B.18) the operator equation

$$L' - L^2 = R(\,\cdot\,, c')c'. \tag{B.19}$$

Recall from equation (B.16) that $B = A'A^{-1}$ may be identified with $-L$ prior to the first focal point to p along c, and recall also the earlier notation of Definition 12.2 of Section 12.1 of $\theta = \text{tr}(B)$. We may then rewrite equation (B.19) as

$$B' = -B \circ B - R(\cdot, c')c' \tag{B.20}$$

in exact accord with equation (12.1) of Section 12.1. Recalling further that $\text{tr}(B') = [\text{tr}(B)]'$, we may obtain either from (B.19) or (B.20) the equation

$$\theta' = -\text{tr}(B \circ B) - \text{Ric}(c', c') = -\text{tr}(L \circ L) - \text{Ric}(c', c') \tag{B.21}$$

which corresponds (with change in sign conventions) to equation (1.5.5) in Karcher (1989): $d/ds[\text{trace}\ (S)] = -\text{trace}(R_N) - \text{trace}(S^2)$, which is as far as Karcher proceeds in decomposing this particular equation. In general relativity, the Raychaudhuri equation for θ' is derived from (B.21) by further decomposing B in terms of the expansion, shear, and vorticity tensors and calculating B^2 in terms of this decomposition (cf. the derivation of equation (12.2) in Section 12.1).

Note also that equation (B.21) occurs in another related form in connection with the "Bochner trick" type identities derived by manipulations of the curvature tensor and by utilizing the fact that the normal field used to calculate the second fundamental tensor is a gradient, i.e., $N = \text{grad}\, f$. In this formalism, taking a parallel orthonormal basis $\{E_1, E_2, \ldots, E_n\}$ along c with $E_n = c'$, one is then led to the following version of (B.21) ($\Box$ denotes the d'Alembertian of f):

$$-\text{Ric}(c', c') = (\Box(f) \circ c)' + \sum_{i,j=1}^{n-1} \left(H^f(E_i, E_j)\right)^2 \tag{B.22}$$

and hence the inequality

$$-\text{Ric}(c', c') \geq (\Box(f) \circ c)' + \frac{1}{n-1}(\Box(f) \circ c)^2, \tag{B.23}$$

which plays a role in the proof of the so-called splitting theorems in both Riemannian and Lorentzian geometry. In Beem, Ehrlich, Markvorsen, and

Galloway (1985, Lemma 4.3 and Corollary 4.5), for instance, the untraced version of equation (B.23) corresponding to equation (B.20) is employed.

Finally, let us conclude this section by deriving as in Karcher (1989) a well-known second link between E_* and the mean curvature $\mathrm{tr}(L(t))$ in the present context by using equations (B.12) and (B.16). Put $a(t) = \det(E_*)$, or more formally, $a(t) = \det(A(t))$. Then using the identity $\mathrm{tr}(A'A^{-1}) = (\det A)^{-1}(\det A)'$ of Section 12.1, Definition 12.2, we obtain

$$a'(t) = \mathrm{tr}(B)\, a(t) = -\mathrm{tr}(L(t))\, a(t) = \theta(A(t))\, a(t). \tag{B.24}$$

Thus the expansion tensor $\theta(A(t))$ is, up to sign, precisely the mean curvature at $c(t)$ of the hypersurface in the family of f-level surfaces passing through the point $c(t)$. Karcher's viewpoint is that the Jacobi equation should be regarded as giving geometric control by considering the Jacobi equation as equivalent to the so-called Riccati inequality (B.23) and to equation (B.24). Of course, equation (B.24) also demonstrates rather explicitly the well-known fact that in the Raychaudhuri framework, the expansion tensor $\theta(A)$ is the quantity that measures the infinitesimal change in hypersurface volume; indeed, in Eschenburg (1975, p. 32) the tensor θ is even called the "volumen-expansion" tensor.

In the context of equation (B.24), the following elementary calculation has been noted. Suppose that $a(t)$ and $b(t)$ are differentiable functions which satisfy the ordinary differential equations

$$a' = f(t)a \tag{B.25}$$

and

$$b' = g(t)b, \tag{B.26}$$

respectively. Then

$$\left(\frac{a}{b}\right)' = (f(t) - g(t))\left(\frac{a}{b}\right). \tag{B.27}$$

Hence if $a(t)$ and $b(t)$ are positive valued and $f \geq g$, then a/b is monotonic nondecreasing. This remark provides, after integrating along radial geodesics emanating from p, a proof of the Bishop–Gromov Volume Comparison Theorem for Riemannian manifolds, choosing $f = d(p, \cdot)$ and taking g to be the corresponding distance function on the appropriately chosen model space of constant curvature [cf. Karcher (1989, pp. 186–187)].

REFERENCES

Abresch, U. (1985), *Lower curvature bounds, Toponogov's theorem and bounded topology*, Ann. Sci. Ec. Norm. Sup. **18**, 651–670.

Abresch, U. (1987), *Lower curvature bounds, Toponogov's theorem and bounded topology II*, Ann. Sci. Ec. Norm. Sup. **20**, 475–502.

Abresch, U., and D. Gromoll (1990), *On complete manifolds with nonnegative Ricci curvature*, Journal of American Math. Soc. **2**, 355–374.

Aiyama, R. (1991), *On complete spacelike surfaces with constant mean curvature in a Lorentzian 3-space form*, Tsukuba J. Math. **15**, 235–247.

Akutagawa, K. (1987), *On spacelike hypersurfaces with constant mean curvature in the de Sitter space*, Math. Zeitschrift **196**, 13–19.

Akutagawa, K., and S. Nishikawa (1990), *The Gauss map and spacelike surfaces with prescribed mean curvature in Minkowski 3–space*, Tohoku Math. J. **42**, 67–82.

Alexandrov, A. (1967), *A contribution to chronogeometry*, Canad. J. Math. **19**, 1119–1128.

Allison, D. (1991), *Pseudo-convexity in Lorentzian doubly warped products*, Geom. Dedicata **39**, 223–227.

Ambrose, W., R. S. Palais, and I. M. Singer (1960), *Sprays*, Anais. Acad. Brasil Cienc. **32**, 163–178.

Anderson, J. L. (1967), *Principles of Relativity Physics*, Academic Press, New York.

Andersson, L., M. Dahl, and R. Howard (1994), *Boundary and lens rigidity of Lorentzian surfaces*, Research Report, Dept. of Math., Univ. of South Carolina **8**.

Andersson, L., and R. Howard (1994), *Comparison and rigidity theorems in semi-Riemannian geometry*, preprint, Dept. of Math., TRITA/MAT–1994–0017, Royal Inst. of Tech., Stockholm.

Ashtekar, A., and A. Magnon-Ashtekar (1979), *Energy momentum in general relativity*, Phys. Review Lett. **43**, 181–184.

Ashtekar, A., and M. Streubel (1981), *Symplectic geometry of radiative modes and conserved quantities at null infinity*, Proc. Roy. Soc. Lond., Ser. A **376**, 585–607.

Auslander, L., and L. Marcus (1959), *Flat Lorentz 3-Manifolds*, Memoir 30, Amer. Math. Soc..

Avez, A. (1963), *Essais de géométrie riemannienne hyperbolique globale. Applicationes à la Relativité Générale*, Ann. Inst. Fourier **132**, 105–190.

Barbance, C. (1980), *Transformations conformes des varietes lorentziennes homogenes*, C. R. Acad. Sci. Paris, Ser. A–B **291**, A347–A350.

Barnet, F. (1989), *On Lie groups that admit left invariant Lorentz metrics of constant sectional curvature*, Illinois J. Math **33**, 631–642.

Barrow, J., G. Galloway, and F. Tipler (1986), *On the closed universe recollapse conjecture*, Mon. Not. Royal Astr. Soc. **223**, 395–405.

Bartnik, R. (1984), *Existence of maximal surfaces in asymptotically flat spacetimes*, Commun. Math. Phys. **94**, 155–175.

Bartnik, R. (1988a), *Regularity of variational maximal surfaces*, Acta Math. **161**, 145–181.

Bartnik, R. (1988b), *Remarks on cosmological space–times and constant mean curvature surfaces*, Comm. Math. Phys. **117**, 615–624.

Beem, J. K. (1976a), *Conformal changes and geodesic completeness*, Commun. Math. Phys. **49**, 179–186.

Beem, J. K. (1976b), *Globally hyperbolic space–times which are timelike Cauchy complete*, Gen. Rel. Grav. **7**, 339–344.

Beem, J. K. (1976c), *Some examples of incomplete space–times*, Gen. Rel. Grav. **7**, 501–509.

Beem, J. K. (1977), *A metric topology for causally continuous completions*, Gen. Rel. Grav. **8**, 245–257.

Beem, J. K. (1978a), *Homothetic maps of the space–time distance function and differentiability*, Gen. Rel. Grav. **9**, 793–799.

Beem, J. K. (1978b), *Proper homothetic maps and fixed points*, Lett. Math. Phys. **2**, 317–320.

Beem, J. K. (1980), *Minkowski space–time is locally extendible*, Commun. Math. Phys. **72**, 273–275.

Beem, J. K. (1994), *Stability of geodesic incompleteness*, in Differential Geometry and Mathematical Physics (J. K. Beem and K. L. Duggal, eds.), Contemporary Math. Series, vol. **170**, American Mathematical Society, pp. 1–12.

Beem, J. K., C. Chicone, and P. E. Ehrlich (1982), *The geodesic flow and sectional curvature of pseudo-Riemannian manifolds*, Geom. Dedicata **12**, 111–118.

Beem, J. K., and P. E. Ehrlich (1977), *Distance lorentzienne finie et géodésiques f-causales incomplètes*, C. R. Acad. Sci. Paris Ser. A **581**, 1129–1131.

Beem, J. K., and P. E. Ehrlich (1978), *Conformal deformations, Ricci curvature and energy conditions on globally hyperbolic space–times*, Math. Proc. Camb. Phil. Soc. **84**, 159–175.

Beem, J. K., and P. E. Ehrlich (1979a), *Singularities, incompleteness and the Lorentzian distance function*, Math. Proc. Camb. Phil. Soc. **85**, 161–178.

Beem, J. K., and P. E. Ehrlich (1979b), *The space–time cut locus*, Gen. Rel. Grav. **11**, 89–103.

Beem, J. K., and P. E. Ehrlich (1979c), *Cut points, conjugate points and Lorentzian comparison theorems*, Math. Proc. Camb. Phil. Soc. **86**, 365–384.

Beem, J. K., and P. E. Ehrlich (1979d), *A Morse index theorem for null geodesics*, Duke Math. J. **46**, 561–569.

Beem, J. K., and P. E. Ehrlich (1981a), *Constructing maximal geodesics in strongly causal space–times*, Math. Proc. Camb. Phil. Soc. **90**, 183–190.

Beem, J. K., and P. E. Ehrlich (1981b), *Stability of geodesic incompleteness for Robertson–Walker space–times*, Gen. Rel. and Grav. **13**, 239–255.

Beem, J. K., and P. E. Ehrlich (1985a), *Incompleteness of timelike submanifolds with nonvanishing second fundamental form*, Gen. Rel. Grav. **17**, 293–300.

Beem, J. K., and P. E. Ehrlich (1985b), *Geodesic completeness of submanifolds in Minkowski space*, Geom. Dedicata **18**, 213–226.

Beem, J. K., and P. E. Ehrlich (1987), *Geodesic completeness and stability*, Math. Proc. Camb. Phil. Soc. **102**, 319–328.

Beem, J. K., P. E. Ehrlich, and S. Markvorsen (1988), *Timelike isometries and Killing fields*, Geom. Dedicata **26**, 247–258.

Beem, J. K., P. E. Ehrlich, S. Markvorsen, and G. Galloway (1984), *A Toponogov splitting theorem for Lorentzian manifolds*, Springer-Verlag Lecture Notes in Math. **1156**, 1–13.

Beem, J. K., P. E. Ehrlich, S. Markvorsen, and G. Galloway (1985), *Decomposition theorems for Lorentzian manifolds with nonpositive curvature*, J. Diff. Geom. **22**, 29–42.

Beem, J. K., P. E. Ehrlich, and T. G. Powell (1982), *Warped product manifolds in relativity*, in Selected Studies: A Volume Dedicated to the Memory of Albert Einstein (T. M. Rassias and G. M. Rassias, eds.), North-Holland, Amsterdam, pp. 41–56.

Beem, J. K., and S. G. Harris (1993a), *The generic condition is generic*, Gen. Rel. and Grav. **25**, 939–962.

Beem, J. K., and S. G. Harris (1993b), *Nongeneric null vectors*, Gen. Rel. and Grav. **25**, 963–973.

Beem, J. K., and P. E. Parker (1984), *Values of pseudo-Riemannian sectional curvature*, Comment. Math. Helvetici **59**, 319–331.

Beem, J. K., and P. E. Parker (1985), *Whitney stability of solvability*, Pac. J. Math. **116**, 11–23.

Beem, J. K., and P. E. Parker (1989), *Pseudoconvexity and geodesic connectedness*, Annali Math. Pura. Appl. **155**, 137–142.

Beem, J. K., and P. E. Parker (1990), *Sectional curvature and tidal accelerations*, J. Math. Phys. **31**, 819–827.

Beem, J. K., and P. Y. Woo (1969), *Doubly Timelike Surfaces*, Memoir 92, Amer. Math. Soc..

Bejancu, A., and K. L. Duggal (1991), *Degenerate hypersurfaces of semi-Riemannian manifolds*, Bull. Inst. Politchnic Iasi **37**, 13–22.

Bejancu, A., and K. L. Duggal (1994), *Lightlike submanifolds of semi-Riemannian manifolds*, Acta Applicandae Mathematicae (to appear).

Benci, V., and D. Fortunato (1990), *Existence of geodesics for the Lorentz metric of a stationary gravitational field*, Ann. Inst. H. Poincaré, Analyse Nonlineare **7**.

Benci, V., and D. Fortunato (1994), *On the existence of infinitely many geodesics on space–time manifolds*, Advances in Math. **105**, 1–25.

Bérard-Bergery, L., and A. Ikemakhen (1993), *On the holonomy of Lorentzian manifolds*, in Differential Geometry: Geometry in Mathematical Physics and Related Topics (R. Greene and S.-T. Yau, eds.), vol. 54, part 2, Amer. Math. Soc. Proceedings of Symposia in Pure Math., pp. 27–40.

Berger, M. (1960), *Les variétés riemanniennes (1/4)-pincées*, Annali della Scuola Normale Sup. di Pisa, Ser. III **14**, 161–170.

Besse, A. L. (1978), *Manifolds all of whose Geodesics are Closed*, Ergebnisse der Mathematik und ihrer Grenzgebiete 93, Springer-Verlag, Berlin.

Besse, A. L. (1987), *Einstein Manifolds*, Ergebnisse der Mathematik und ihrer Grenzgebiete, 3 Folge, Band 10, Springer-Verlag, Berlin.

Birkhoff, G., and G.-C. Rota (1969), *Ordinary Differential Equations, 2nd ed.*, Blaisdell, Waltham, Massachusetts.

Birman, G., and K. Nomizu (1984), *The Gauss–Bonnet theorem for two-dimensional spacetimes*, Mich. Math. J. **31**, 77–81.

Bishop, R. L., and R. Crittenden (1964), *Geometry of Manifolds*, Academic Press, New York.

Bishop, R. L., and B. O'Neill (1969), *Manifolds of negative curvature*, Trans. Amer. Math. Soc. **145**, 1–49.

Bölts, G. (1977), *Existenz und Bedeutung von konjugierten Werten in der Raum-Zeit*, Bonn Universität Diplomarbeit, Bonn.

Bondi, H. (1960), *Gravitational waves in general relativity*, Nature **186**, 535.

Bondi, H. (1968), *Cosmology, 2nd ed.*, Cambridge University Press, Cambridge.

Bondi, H., F. Pirani, and I. Robinson (1959), *Gravitational waves in general relativity III; Exact plane waves*, Proc. Roy. Soc. London A **251**, 519–533.

Bonnor, W. (1972), *Null hypersurfaces in Minkowski space–time*, Tensor N. S. **24**, 329–345.

Boothby, W. M. (1986), *An Introduction to Differentiable Manifolds and Riemannian Geometry, 2nd ed.*, Pure and Applied Math. Ser., vol. 120, Academic Press, New York.

Borde, A. (1985), *Singularities in closed spacetimes*, Class. Quantum Grav. **2**, 589–596.

Bosshard, B. (1976), *On the b-boundary of the closed Friedmann model*, Commun. Math. Phys. **46**, 263–268.

Boyer, R. H., and R. W. Lindquist (1967), *Maximal analytic extension of the Kerr metric*, J. Math. Phys. **8**, 265–281.

Boyer, R. H., and T. G. Price (1965), *An interpretation of the Kerr metric in General Relativity*, Proc. Camb. Phil. Soc. **61**, 531–534.

Brill, D., and F. Flaherty (1976), *Isolated maximal surfaces in spacetime*, Commun. Math. Phys. **50**, 157–165.

Brinkmann, H. (1925), *Einstein spaces which are mapped on each other*, Math. Ann. **94**, 119–145.

Budic, R., J. Isenberg, L. Lindblom, and P. Yasskin (1978), *On the determination of Cauchy surfaces from intrinsic properties*, Comm. Math. Phys. **61**, 87–95.

Budic, R., and R. K. Sachs (1974), *Causal boundaries for general relativistic space-times*, J. Math. Phys. **15**, 1302–1309.

Budic, R., and R. K. Sachs (1976), *Scalar time functions: differentiability*, in Differential Geometry and Relativity (M. Cohen and M. Flato, eds.), Reidel, Dordrecht, pp. 215–224.

Burns, J. (1977), *Curvature functions on Lorentz 2-manifolds*, Pacific J. Math. **70**, 325–335.

Busemann, H. (1932), *Über die Geometrien, in denen die "Kreise mit unendlichem Radius" die kürzesten Linien sind*, Math. Annalen **106**, 140–160.

Busemann, H. (1942), *Metric Methods in Finsler Spaces and in the Foundations of Geometry*, Annals of Math. Studies, vol. 8, Princeton University Press, Princeton, New Jersey.

Busemann, H. (1955), *The Geometry of Geodesics*, Academic Press, New York.

Busemann, H. (1967), *Timelike Spaces*, Dissertationes Math. Rozprawy Mat. **53**.

Busemann, H., and J. K. Beem (1966), *Axioms for indefinite metrics*, Rnd. Cir. Math. Palermo **15**, 223–246.

Cahen, M., J. Leroy, M. Parker, F. Tricerri, and L. Vanhecke (1990), *Lorentz manifolds modelled on a Lorentz symmetric space*, J. Geom Phys. **7**, 571–581.

Cahen, M., and M. Parker (1980), *Pseudo-riemannian Symmetric Spaces*, Memoir 229, Amer. Math. Soc..

Calabi, E. (1957), *An extension of E. Hopf's maximum principle with an application to Riemannian geometry*, Duke Math. J. **25**, 45–56.

Calabi, E. (1968), *Examples of Bernstein problems for some nonlinear equations*, Proc. Symp. Pure and Applied Math. **15**, 223–230.

Canarutto, D., and C. Dodson (1985), *On the bundle of principal connections and the stability of b-incompleteness of manifolds*, Math. Proc. Camb. Phil. Soc. **98**, 51–59.

Canarutto, D., and P. Michor (1987), *On the stability of b-incompleteness in the Whitney topology on the space of connections*, Istit. Lombardo Acad. Sci. Lett. Rend. A **121**, 217–226.

Carafora, M., and A. Marzuoli (1987), *Smoothing out spacetime geometry*, Seventh Italian Conference on General Relativity and Gravitational Physics, (Rapallo, 1986), World Sci. Publishing, Singapore, 19–34.

Carot, J., and J. da Costa (1993), *On the geometry of warped space–times*, Class. Quantum Grav. **10**, 461–482.

Carrière, Y. (1989), *Autour de la Conjecture de L. Markus sur les variétes affines*, Invent. Math. **95**, 615–628.

Carter, B. (1971a), *Causal structure in space–time*, Gen. Rel. Grav. **1**, 349–391.

Carter, B. (1971b), *Axisymmetric black hole has only two degrees of freedom*, Phys. Rev. Lett. **26**, 331–333.

Chandrasekhar, S. (1983), *The Mathematical Theory of Black Holes*, Oxford University Press.

Cheeger, J., and D. Ebin (1975), *Comparison Theorems in Riemannian Geometry*, North-Holland, Amsterdam.

Cheeger, J., and D. Gromoll (1971), *The splitting theorem for manifolds of nonnegative Ricci curvature*, J. Diff. Geo. **6**, 119–128.

Cheeger, J., and D. Gromoll (1972), *On the structure of complete manifolds of nonnegative curvature*, Ann. Math. **96**, 413–433.

Cheng, S.-Y., and S.-T. Yau (1976), *Maximal spacelike hypersurfaces in Lorentz–Minkowski space*, Ann. of Math. **104**, 407–419.

Chicone, C., and P. Ehrlich (1980), *Line integration of Ricci curvature and conjugate points in Lorentzian and Riemannian manifolds*, Manuscripta Math. **31**, 297–316.

Chicone, C., and P. E. Ehrlich (1984), *Gradient-like and integrable vector fields on* $\mathbb{R}^2$, Manuscripta Math. **49**, 141–164.

Chicone, C., and P. E. Ehrlich (1985), *Lorentzian geodesibility*, in Commemorative Volume in Differential Topology, Geometry and Related Fields, and their Applications to the Physical Sciences and Engineering (G. Rassias, ed.), Teubner Text zur Mathematik, pp. 75–99.

Choi, H., and A. Treibergs (1990), *Gauss map of spacelike constant mean curvature hypersurface of Minkowski space*, J. Diff. Geom. **32**, 775–817.

Choquet–Bruhat, Y., and C. DeWitt–Morette (1989), *Analysis, Manifolds and Physics Part II: 92 Applications*, North-Holland, Amsterdam.

Choquet–Bruhat, Y., C. DeWitt–Morette, and M. Dillard–Bleick (1982), *Analysis, Manifolds and Physics, Revised Edition*, North-Holland, Amsterdam.

Choquet–Bruhat, Y., A. E. Fischer, and J. E. Marsden (1979), *Maximal hypersurfaces and positivity of mass*, in Proceedings of 1976 International Summer School of Italian Physical Society, "Enrico Fermi," Course LXVII, Isolated Gravitating Systems in General Relativity (J. Ehlers, ed.), North Holland, New York, pp. 396–456.

Choquet–Bruhat, Y., and R. Geroch (1969), *Global aspect of the Cauchy problem*, Commun. Math. Phys. **14**, 329–335.

Christodoulou, D., and S. Klainerman (1993), *The Global Nonlinear Stability of Minkowski Space*, Princeton University Press, Princeton.

Chrusciel, P. T. (1991), *On Uniqueness in the Large of Solutions of Einstein Equations ("Strong Cosmic Censorship")*, Australian University Press, Canberra.

Ciufolini, I., and J. A. Wheeler (1995), *Gravitation and Inertia*, Princeton Series in Physics, Princeton University Press.

Clarke, C. J. S. (1970), *On the global isometric embedding of pseudo-Riemannian manifolds*, Proc. Roy. Soc. Lond. **A314**, 417–428.

Clarke, C. J. S. (1971), *On the geodesic completeness of causal space–times*, Proc. Camb. Phil. Soc. **69**, 319–324.

Clarke, C. J. S. (1973), *Local extensions in singular space–times*, Commun. Math. Phys. **32**, 205–214.

Clarke, C. J. S. (1975), *Singularities in globally hyperbolic space–times*, Commun. Math. Phys. **41**, 65–78.

Clarke, C. J. S. (1976), *Space–time singularities*, Commun. Math. Phys. **49**, 17–23.

Clarke, C. J. S. (1982), *Space–times of low differentiability and singularities*, J. Math. Anal. Appl. **88**, 270–305.

Clarke, C. J. S. (1993), *The Analysis of Space–time Singularities*, Cambridge Lecture Notes in Physics, vol. 1, Cambridge University Press.

Clarke, C. J. S., and B. G. Schmidt (1977), *Singularities: the state of the art*, Gen. Rel. and Grav. **8**, 129–137.

Cohn–Vossen, S. (1936), *Total Krümmung und geodatische linien auf einfach zusammenhängenden offenen vollständigen Flächenstücken,*, Recueil Mathematique **1**, 139–163.

Coll, B., and J.-A. Morales (1988), *Sur les reperes symetriques lorentziens*, C. R. Acad. Sci. Paris **306**, 791–794.

Conlon, L. (1993), *Differentiable Manifolds: a first course*, Birkhauser, Boston.

Cordero, L. A., and P. E. Parker (1995a), *Examples of sectional curvature with prescribed symmetry on 3-manifolds*, Czech. Math. J. **45**, 7–20.

Cordero, L. A., and P. E. Parker (1995b), *Left-invariant Lorentzian metrics on 3-dimensional Lie groups*, Rend. Math. Appl. (to appear).

Cordero, L. A., and P. E. Parker (1995c), *Symmetries of sectional curvature on 3-manifolds*, Demonstratio Math. **28** (to appear).

Cormack, W., and G. Hall (1979), *Riemannian curvature and the classification of the Riemann and Ricci tensors in space–times*, Internat. J. Theoretic. Phys. **18**, 279–289.

Crittenden, R. (1962), *Minimum and conjugate points in symmetric spaces*, Canad. J. Math. **14**, 320–328.

Dajczer, M., and K. Nomizu (1980a), *On the boundedness of Ricci curvature of an indefinite metric*, Bol. Soc. Brazil Math. **11**, 25–30.

Dajczer, M., and K. Nomizu (1980b), *On sectional curvature of indefinite metrics II*, Math. Annalen **247**, 279–282.

Dederzinski, A. (1993), *Geometry of Elementary Particles*, in Differential Geometry: Geometry in Mathematical Physics and Related Topics (R. Greene and S.-T. Yau, eds.), Amer. Math. Soc. Proceedings of Symposia in Pure Math., 54, part 2, pp. 157–171.

De Felice, F., and C. J. S. Clarke (1990), *Relativity on Curved Manifolds*, Cambridge Monographs on Mathematical Physics, Cambridge University Press.

Del Riego, L. (1993), *Spaces of Sprays*, in Taller De Geometria Diferencial Sobre Espacios de Geometria (L. Del Riego and C. T. J. Dodson, eds.), vol. 8, Sociedad Matematica Mexicana Aportaciones Matematicas Notas de Investigacion, pp. 81–96.

Del Riego, L., and C. T. J. Dodson (1988), *Sprays, universality and stability*, Math. Proc. Camb. Phil. Soc. **103**, 515–534.

Del Riego, L., and P. E. Parker (1995), *Pseudoconvex and disprisoning homogeneous sprays*, Geom. Dedicata **55**, 211–220.

Dieckmann, J. (1987), *Volumenfunktionen in der allgemeinen Relativitätstheorie*, Dissertation, Technical University Berlin, Fachbereich Mathematik D83.

Dieckmann, J. (1988), *Volume functions in General Relativity*, Gen. Rel. Grav. **20**, 859–867.

Dillen, F., R. Rosca, L. Verstraelen, and L. Vrancken (1992), *Pseudo-isotropic Lorentzian hypersurfaces in Minkowski space*, J. Geom. Phys. **9**, 149–154.

Do Carmo, M. P. (1976), *Differential Geometry of Curves and Surfaces*, Prentice Hall, Englewood Cliffs, New Jersey.

Dodson, C. T. J. (1978), *Space–time edge geometry*, Int. J. Theor. Phys. **17**, 389–504.

Dodson, C. T. J. (1980), *Categories, Bundles and Spacetime Topology*, Shiva Math. Series 1, Shiva, Kent.

Dodson, C. T. J., and T. Poston (1977), *Tensor Geometry: The Geometric Viewpoint and Its Uses*, Survey and Reference Works in Math. 1, Pitman, San Francisco.

Dodson, C. T. J., and L. J. Sulley (1980), *On bundle completion of parallelizable manifolds*, Math. Proc. Camb. Phil. Soc. **87**, 523–525.

Domiaty, R. (1985), *On bijections of Lorentz manifolds which leave the class of spacelike paths invariant*, Topology Appl. **20**, 39–46.

Drumm, T. (1993), *Margulis space–times*, in Differential Geometry: Geometry in Mathematical Physics and Related Topics (R. Greene and S.-T. Yau, eds.), Amer. Math. Soc. Proceedings of Symp. Pure Math., 54, Part 2, pp. 191–195.

Drumm, T., and W. Goldman (1990), *Complete flat Lorentz 3-manifolds with free fundamental group*, Internat. J. Math. **1**, 149–161.

Duggal, K. L. (1990), *Lorentzian geometry of globally framed manifolds*, Acta Applicandae Math. **19**, 131–148.

Duggal, K. L., and A. Bejancu (1992), *Lightlike submanifolds of codimension two*, Math. J. Toyama Univ. **15**, 59–82.

Duggal, K. L., and A. Bejancu (1993), *Light-like CR-hypersurfaces of indefinite Kähler manifolds*, Acta Applicandae Mathematicae **31**, 171–190.

Eardley, D. M., J. Isenberg, J. Marsden, and V. Moncrief (1986), *Homothetic and conformal symmetries of solutions to Einstein's equations*, Commun. Math. Phys. **106**, 137–158.

Eardley, D. M., and L. Smarr (1979), *Time functions in numerical relativity: Marginally bound dust collapse*, Phys. Rev. D **19**, 2239–2259.

Easley, K. L. (1991), *Local existence of warped product metrics*, Ph.D. Thesis, University of Missouri–Columbia.

Eberlein, P. (1972), *Product manifolds that are not negative space forms*, Mich. Math. J. **19**, 225–231.

Eberlein, P. (1973a), *Geodesic flows on negatively curved manifolds, II*, Trans. Amer. Math. Soc. **178**, 57–82.

Eberlein, P. (1973b), *When is a geodesic flow of Anosov type I*, J. Diff. Geo. **8**, 437–463.

Eberlein, P., and B. O'Neill (1973), *Visibility manifolds*, Pacific J. Math. **46**, 45–109.

Ehlers, J., and W. Kundt (1962), *Exact solutions of the gravitational field equations*, in Gravitation (L. Witten, ed.), Wiley, Chichester.

Ehresmann, C. (1951), *Les connexions infinitésimales dans un espace fibré différentiable*, Colloque de Topologie (Espaces Fibrés), Bruxelles 1950, Masson, Paris, 29–55.

Ehrlich, P. E. (1974), *Continuity properties of the injectivity radius function*, Compositio Math. **29**, 151–178.

Ehrlich, P. E. (1976a), *Metric deformations of curvature I: local convex deformations*, Geometriae Dedicata **5**, 1–24.

Ehrlich, P. E. (1976b), *Metric deformations of curvature II: compact 3-manifolds*, Geometriae Dedicata **5**, 147–161.

Ehrlich, P. E. (1982), *The displacement function of a timelike isometry*, Tensor, N. S. **38**, 29–36.

Ehrlich, P. E. (1991), *Null cones and pseudo-Riemannian metrics*, Semigroup Forum **43**, 337–343.

Ehrlich, P. E. (1993), *Astigmatic conjugacy and achronal boundaries*, in Geometry and Global Analysis (T. Kotake, S. Nishikawa, and R. Schoen, eds.), Report of the First MSJ International Research Institute, July 12–23, 1993, Tohoku University, Sendai, Japan, pp. 197–208.

Ehrlich, P. E., and G. Emch (1992a), *Gravitational waves and causality*, Reviews in Mathematical Physics **4**, 163–221, Errata 4, 501.

Ehrlich, P. E., and G. Emch (1992b), *Quasi-time functions in Lorentzian geometry*, Marcel Dekker Lecture Notes in Pure and Applied Mathematics **144**, 203–212.

Ehrlich, P. E., and G. Emch (1992c), *The conjugacy index and simple astigmatic focusing*, Contemporary Mathematics **127**, 27–39.

Ehrlich, P. E., and G. Emch (1993), *Geodesic and causal behavior of gravitational plane waves: astigmatic conjugacy*, in Proc. Symposia in Pure Mathematics, Amer. Math. Soc., Vol. 54, Part 2, pp. 203–209.

Ehrlich, P. E., and G. Galloway (1990), *Timelike lines*, Class. Quantum Grav. **7**, 297–307.

Ehrlich, P. E., and S.-B. Kim (1989a), *A Morse index theorem for null geodesics with spacelike endmanifolds*, a volume in Geometry and Topology (G. Stratopoulos and G. Rassias, ed.), World Scientific Publishing Co., pp. 105–133.

Ehrlich, P. E., and S.-B. Kim (1989b), *A focal index theorem for null geodesics*, J. Geom. and Physics **6**, 657–670.

Ehrlich, P. E., and S.-B. Kim (1991), *A focal Rauch comparison theorem for null geodesics*, Commun. Korean Math. Soc. **6**, 73–87.

Ehrlich, P. E., and S.-B. Kim (1994), *From the Riccati inequality to the Raychaudhuri equation*, in Differential Geometry and Mathematical Physics (J. K. Beem and K. L. Duggal, eds.), Contemporary Math. Series, vol. **170**, American Mathematical Society, pp. 65–78.

Einstein, A. (1916), *Die Grundlage der allgemeinen Relativitätstheorie*, Annalen der Phys. **49**, 769–822.

Einstein, A. (1953), *The Meaning of Relativity, 4th ed.*, Princeton University Press, Princeton, New Jersey.

Einstein, A. and N. Rosen (1937), *On Gravitational Waves*, Jour. Franklin Institute **223**, 43–54.

Ellis, G. F. R., and B. G. Schmidt (1977), *Singular space–times*, Gen. Rel. Grav. **8**, 915–953.

Emch, G. (1984), *Mathematical and conceptual foundations of 20th century physics*, North-Holland, Amsterdam.

Eschenburg, J.-H. (1975), *Stabilitätsverhalten des Geodätischen Flüsses Riemannscher Mannigfaltigkeiten*, Thesis, Bonn University, also Bonner Math. Schr. vol. 87 (1976).

Eschenburg, J.-H. (1987), *Comparison theorems and hypersurfaces*, Manuscripta Math. **59**, 295–323.

Eschenburg, J.-H. (1988), *The splitting theorem for space–times with strong energy condition*, J. Diff. Geom. **27**, 477–491.

Eschenburg, J.-H. (1989), *Maximum principal for hypersurfaces*, Manuscripta Math. **64**, 55–75.

Eschenburg, J.-H., and G. Galloway (1992), *Lines in space–times*, Commun. Math. Phys. **148**, 209–216.

Eschenburg, J.-H., and E. Heintze (1984), *An elementary proof of the Cheeger–Gromoll splitting theorem*, Ann. Global Analysis Geometry **2**, 141–151.

Eschenburg, J.-H., and J. O'Sullivan (1976), *Growth of Jacobi fields and divergence of geodesics*, Math. Zeitschrift **150**, 221–237.

Eschenburg, J.-H., and J. O'Sullivan (1980), *Jacobi tensors and Ricci curvature*, Math. Annalen **252**, 1–26.

Everson, J., and C. J. Talbot (1976), *Morse theory on timelike and causal curves*, Gen. Rel. Grav. **7**, 609–622.

Everson, J., and C. J. Talbot (1978), *Erratum: Morse theory on timelike and causal curves*, Gen. Rel. Grav. **9**, 1047.

Fama, C. J., and S. M. Scott (1994), *Invariance properties of boundary sets of open embeddings of manifolds and their application to the abstract boundary*, in Differential Geometry and Mathematical Physics (J. K. Beem and K. L. Duggal, eds.), Contemporary Math. Series, vol. **170**, American Mathematical Society, pp. 79–111.

Fegan, H., and R. Millman (1978), *Quadrants of Riemannian metrics*, Mich. Math. J. **25**, 3–7.

Fialkow, A. (1938), *Hypersurfaces of a space of constant curvature*, Ann. of Math. **39**, 762–785.

Fierz, M., and J. Jost (1965), *Affine vollstandigkeit und kompakte Lorentzische mannigfaltigkeiten*, Helv. Phys. Acta **38**, 137–141.

Fischer, A. E., and J. E. Marsden (1972), *The Einstein equations of evolution—a geometric approach*, J. Math. Phys. **13**, 546–568.

Flaherty, F. (1975a), *Lorentzian manifolds of nonpositive curvature*, Proc. Symp. Pure Math. **27**, no. 2, Amer. Math. Soc., 395–399.

Flaherty, F. (1975b), *Lorentzian manifolds of nonpositive curvature II*, Proc. Amer. Math. Soc. **48**, 199–202.

Flaherty, F. (1979), *The boundary value problem for maximal hypersurfaces*, Proc. Nat. Acad. Sci. U.S.A. **76**, 4765–4767.

Fock, V. (1966), *The Theory of Space, Time and Gravitation, 2nd Revised Edition*, Pergamon Press, Oxford.

Frankel, T. (1979), *Gravitational Curvature*, W. H. Freeman, San Francisco.

Frankel, T., and G. Galloway (1981), *Energy density and spatial curvature in general relativity*, J. Math. Phys. **22**, 813–817.

Frankel, T., and G. Galloway (1982), *Stable minimal surfaces and spatial topology in general relativity*, Math. Zeit. **181**, 395–406.

Freudenthal, H. (1931), *Über die enden topologischer Raüme und Gruppen*, Math. Zeitschrift **33**, 692–713.

Friedlander, F. G. (1975), *The Wave Equation on a Curved Space–time*, Cambridge University Press, Cambridge.

Galloway, G. (1977), *Closure in anisotropic cosmological models*, J. Math. Phys. **18**, 250–252.

Galloway, G. (1979), *A generalization of Myers' Theorem and an application to relativistic cosmology*, J. Diff. Geo. **14**, 105–116.

Galloway, G. (1980), *On the topology of Wheeler universes*, Phys. Lett. **79A**, 369–370.

Galloway, G. (1982), *Some global properties of closed spatially homogeneous space–times*, Gen. Rel. Grav. **14**, 87–96.

Galloway, G. (1983), *Causality violation in spatially closed space–times*, Gen. Rel. Grav. **15**, 165–171.

Galloway, G. (1984a), *Some global aspects of compact space–times*, Arch. Math. **42**, 168–172.

Galloway, G. (1984b), *Closed timelike geodesics*, Trans. Amer. Math. Soc. **285**, 379–388.

Galloway, G. (1984c), *Splitting theorems for spatially closed space–times*, Commun. Math. Phys. **96**, 423–429.

Galloway, G. (1985), *Some results on Cauchy surface criteria in Lorentzian geometry*, Illinois Math. J. **29**, 1–10.

Galloway, G. (1986a), *Curvature, causality and completeness in space–times with causally complete spacelike slices*, Math. Proc. Camb. Phil. Soc. **99**, 367–375.

Galloway, G. (1986b), *Compact Lorentzian manifolds without closed nonspacelike geodesics*, Proc. Amer. Math. Soc. **98**, 119–124.

Galloway, G. (1989a), *The Lorentzian splitting theorem without completeness assumption*, J. Diff. Geom. **29**, 373–387.

Galloway, G. (1989b), *Some connections between global hyperbolicity and geodesic completeness*, J. Geom. and Phys. **6**, 127–141.

Galloway, G. (1993), *The Lorentzian version of the Cheeger–Gromoll splitting theorem and its applications to General Relativity*, in Differential Geometry: Geometry in Mathematical Physics and Related Topics (R. Greene and S.-T. Yau, eds.), vol. 54, part 2, Amer. Math. Soc. Proceedings of Symposia in Pure Math.,, pp. 249–257.

Galloway, G., and A. Horta (1995), *Regularity of Lorentzian Busemann functions*, Trans. Amer. Math. Soc. (to appear).

Garcia–Rio, E., and D. Kupeli (1995), *Singularity versus splitting theorems for stably causal space–times*, preprint, April, 1995, Facultade de Matematicas, Universidade de Santiago de Compostela, Santiago, Spain.

Garfinkle, D., and Q. Tian (1987), *Spacetimes with cosmological constant and a conformal Killing field have constant curvature*, Class. Quantum Grav. **4**, 137–139.

Gerhardt, C. (1983), *Maximal H-surfaces in Lorentzian manifolds*, Commun. Math. Phys. **96**, 523–553.

Geroch, R. P. (1966), *Singularities in closed universes*, Phys. Review Letters **17**, 445–447.

Geroch, R. P. (1968a), *Spinor structure of space–times in general relativity I*, J. Math. Phys. **9**, 1739–1744.

Geroch, R. P. (1968b), *What is a singularity in general relativity*, Ann. Phys. (N.Y.) **48**, 526–540.

Geroch, R. P. (1969), *Limits of spacetimes*, Commun. Math. Phys. **13**, 180–193.

Geroch, R. P. (1970a), *Domain of dependence*, J. Math. Phys. **11**, 437–449.

Geroch, R. P. (1970b), *Singularities in Relativity*, in Relativity (M. Carmeli, S. Fickler, and L. Witten, eds.), Plenum, New York, pp. 259–291.

Geroch, R. P., and G. T Horowitz (1979), *Global structure of space–time*, in General Relativity: An Einstein Centenary Survey (S. Hawking and W. Israel, eds.), Cambridge University Press, Cambridge, pp. 212–293.

Geroch, R. P., E. H. Kronheimer, and R. Penrose (1972), *Ideal points in space–time*, Proc. Roy. Soc. Lond. **A327**, 545–567.

Geroch, R. P., and J. Winicour (1981), *Linkages in general relativity*, J. Math. Phys. **22**, 803–812.

Göbel, R. (1976), *Zeeman topologies on space–times of general relativity theory*, Commun. Math. Phys **46**, 289–307.

Göbel, R. (1980), *Natural topologies on Lorentzian manifolds*, Mitt. Math. Ges. Hamburg **10**, 763–771.

Goddard, A. (1977a), *Foliations of space–times by spacelike hypersurfaces of constant mean curvature*, Comm. Math. Phys. **54**, 279–282.

Goddard, A. (1977b), *Some remarks on the existence of spacelike hypersurfaces of constant mean curvature*, Math. Proc. Camb. Phil. Soc. **82**, 489–495.

Goldman, W., and Y. Kamishima (1984), *The fundamental group of a compact flat Lorentz space form is virtually polycyclic*, J. Diff. Geom **19**, 233–240.

Graves, L. (1979), *Codimension one isometric immersions between Lorentz spaces*, Trans. Amer. Math. Soc. **252**, 367–392.

Graves, L., and K. Nomizu (1978), *On sectional curvature of indefinite metrics*, Math. Ann. **232**, 267–272.

Gray, A. (1990), *Tubes*, Addison-Wesley, Reading, Massachusetts.

Green, L. W. (1958), *A theorem of E. Hopf*, Mich. Math. J. **5**, 31–34.

Greene, R. E. (1970), *Isometric Embeddings of Riemannian and Pseudo-Riemannian Manifolds*, Memoir 97, Amer. Math. Soc..

Gromoll, D., W. Klingenberg, and W. Meyer (1975), *Riemannsche Geometrie im Grossen*, Lecture Notes in Mathematics, vol. 55, Springer-Verlag, Berlin.

Gromoll, D., and W. Meyer (1969), *On complete open manifolds of positive curvature*, Ann. of Math. **90**, 75–90.

Guimares, F. (1992), *The integral of the scalar curvature of complete manifolds without conjugate points*, J. Diff. Geom. **32**, 651–662.

Gulliver, R. (1975), *On the variety of manifolds without conjugate points*, Trans. Amer. Math. Soc. **210**, 185–201.

Hall, G. S. (1983), *Curvature collineations and determination of the metric from the curvature*, Gen. Rel. and Grav. **15**, 581–589.

Hall, G. S. (1984), *The significance of curvature in general relativity*, Gen. Rel. and Grav. **16**, 495–500.

Hall, G. S. (1987), *Curvature, metric and holonomy in general relativity*, in Differential geometry and its applications, Univ. J. E. Purkyne, Brno, pp. 127–136.

Hall, G. S. (1992), *Weyl manifolds and connections*, J. Math. Phys. **33**, 2633–2638.

Hall, G. S., and D. Hossack (1993), *Some remarks on sectional curvature and tidal accelerations*, J. Math. Phys. **34**, 5897–5899.

Hall, G. S., and W. Kay (1988), *Curvature structure in general relativity*, J. Math. Phys. **29**, 420–427.

Hano, J., and K. Nomizu (1983), *On isometric immersions of the hyperbolic plane into the Lorentz–Minkowski space and the Monge–Ampere equations of a certain type*, Math. Ann. **262**, 245–253.

Harris, S. G. (1979), *Some comparison theorems in the geometry of Lorentz manifolds*, Ph.D. Thesis, University of Chicago.

Harris, S. G. (1982a), *A triangle comparison theorem for Lorentz manifolds*, Indiana Math. J. **31**, 289–308.

Harris, S. G. (1982b), *On maximal geodesic diameter and causality in Lorentzian manifolds*, Math. Ann. **261**, 307–313.

Harris, S. G. (1985), *A characterization of Robertson–Walker spaces by null sectional curvature*, Gen. Rel. Grav. **17**, 493–498.

Harris, S. G. (1987), *Complete codimension-one spacelike immersions*, Class. Quantum Grav. **4**, 1577–1585.

Harris, S. G. (1988a), *Closed and complete spacelike hypersurfaces in Minkowski space*, Class. Quantum Grav. **5**, 111–119.

Harris, S. G. (1988b), *Complete spacelike immersions with topology*, Class. Quantum Grav. **5**, 833–838.

Harris, S. G. (1992), *Conformally stationary space–times*, Class. Quantum Grav. **9**, 1823–1827.

Harris, S. G. (1993), *What is the shape of space in a spacetime?*, in Differential Geometry: Geometry in Mathematical Physics and Related Topics (R. Greene and S.-T. Yau, eds.), vol. 54, part 2, Amer. Math. Soc. Proceedings of Symposia in Pure Mathematics,, pp. 287–296.

Harris, S. G. (1994), *The method of timelike 2-planes*, in Differential Geometry and Mathematical Physics (J. K. Beem and K. L. Duggal, eds.), Contemporary Math. Series vol. **170**, Amer. Math. Soc., pp. 125–134.

Hartman, P. (1964), *Ordinary Differential Equations*, Wiley, New York.

Hawking, S. W. (1967), *The occurrence of singularities in cosmology III: Causality and singularities.*, Proc. Roy. Soc. Lond. **A300**, 187–201.

Hawking, S. W. (1968), *The existence of cosmic time functions*, Proc. Roy. Soc. Lond. **A308**, 433–435.

Hawking, S. W. (1971), *Stable and generic properties in general relativity*, Gen. Rel. Grav. **1**, 393–400.

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

Hawking, S. W., A. R. King, and P. J. McCarthy (1976), *A new topology for curved space–time which incorporates the causal, differential and conformal structures*, J. Math. Phys. **17**, 174–181.

Hawking, S. W., and R. Penrose (1970), *The singularities of gravitational collapse and cosmology*, Proc. Roy. Soc. Lond. **A314**, 529–548.

Hawking, S. W., and R. K. Sachs (1974), *Causally continuous space–times*, Commun. Math. Phys. **35**, 287–296.

Helfer, A. (1990), *The angular momentum of gravitational radiation*, Phys. Lett. A **150**, 342–344.

Helfer, A. (1992), *Difficulties with quasilocal momentum and angular momentum*, Class. Quantum Grav. **9**, 1001–1008.

Helfer, A. (1993), *The kinematics of the gravitational field*, in Differential Geometry: Geometry in Mathematical Physics and Related Topics (R. Greene and S.-T, Yau, eds.), vol. 54, part 2, Amer. Math. Soc. Proceedings of Symp. in Pure Mathematics, pp. 297–316.

Helfer, A. D. (1994a), *Conjugate points and higher Arnol'd–Maslov classes*, in Differential Geometry and Mathematical Physics (J. K. Beem and K. L. Duggal, eds.), Contemporary Math. Series, vol. **170**, American Mathematical Society, pp. 135–148.

Helfer, A. D. (1994b), *Conjugate points on spacelike geodesics or pseudo-self-adjoint Morse–Sturm–Liouville systems*, Pacific J. of Math. **164**, 321–350.

Helgason, S. (1978), *Differential Geometry, Lie Groups and Symmetric Spaces*, Academic Press, New York.

Hermann, R. (1968), *Differential Geometry and the Calculus of Variations*, Academic Press, New York.

Hicks, N. J. (1965), *Notes on Differential Geometry*, D. Van Nostrand, Princeton, New Jersey.

Hilgert, J., K. Hofmann, and J. Lawson (1989), *Lie Groups, Convex Cones, and Semigroups*, Oxford Univ. Press, Oxford.

Hirsch, M. (1976), *Differential Topology*, Grad. Texts in Math., vol. 33, Springer-Verlag, New York.

Hopf, E. (1948), *Closed surfaces without conjugate points*, Proc. Nat. Acad. Sci. **34**, 47–51.

Hopf, H., and W. Rinow (1931), *Über den Begriff des vollständigen differentialgeometrischen Fläche*, Comment. Math. Helv. **3**, 209–225.

Ihrig, E. (1975), *An exact determination of the gravitational potentials g_{ij} in terms of the gravitational fields ${R^l}_{ijk}$*, J. Math. Phys. **16**, 54–55.

Ikawa, T., and H. Nakagawa (1988), *A remark on totally vicious space–times*, J. of Geometry **32**, 51–54.

Israel. W. (1994), *The internal geometry of black holes*, in Differential Geometry and Mathematical Physics (J. K. Beem and K. L. Duggal, eds.), Contemporary Math. Series, vol **170**, Amer. Math. Soc., pp. 125–134.

Jee, D.-J. (1984), *Gauss–Bonnet formula for general Lorentzian surfaces*, Geom. Dedicata **15**, 215–231.

Johnson, R. A. (1977), *The bundle boundary in some special cases*, J. Math. Phys. **18**, 898–902.

Kannar, J., and I. Racz (1992), *On the strength of space–time singularities*, J. Math. Phys. **33**, 2842–2848.

Karcher, H. (1982), *Infinitesmiale charakterisierung von Friedmann–Universen*, Archiv. Math. **38**, 58–64.

Karcher, H. (1989), *Riemannian comparison constructions*, in Global Differential Geometry (S. S. Chern, ed.), M.A.A. Studies in Mathematics, vol. 27, pp. 170–222.

Katsuno, K. (1980), *Null hypersurfaces in Lorentzian manifolds I*, Math. Proc. Camb. Phil. Soc. **88**, 175–182.

Kelley, J. L. (1955), *General Topology*, Univ. Ser. in Higher Math., D. Van Nostrand, Princeton, New Jersey.

Kelly, R., K. Tod, and N. Woodhouse (1986), *Quasi-local mass for small surfaces*, Class. Quantum Grav. **3**, 1151–1167.

Kerchove, M. (1991), *The structure of Einstein spaces admitting conformal motions*, Class. Quantum Grav. **8**, 819–825.

Ki, U.-H., H.-J. Kim, and H. Nakagawa (1991), *On spacelike hypersurfaces with constant mean curvature of a Lorentz space form*, Tokyo J. Math. **14**, 205–216.

Kim, J.-C., and J.-H. Kim (1986), *Wave Surfaces*, Bull. Korean Math. Soc. **23**, 31–33.

Klingenberg, W. (1959), *Contributions to Riemannian geometry in the large*, Ann. of Math. **69**, 654–666.

Klingenberg, W. (1961), *Über Riemannsche Mannigfaltigkeiten mit positiver Krümmung*, Comment. Math. Helv. **35**, 47–54.

Klingenberg, W. (1962), *Über Riemannsche Mannigfaltigkeiten mit nach oben beschränkter Krümmung*, Annali di Mat. **60**, 49–59.

Klingenberg, W. (1978), *Lectures on Closed Geodesics*, Grundlehren der mathematischen Wissenschaften, vol. 230, Springer-Verlag, New York.

Ko, M., E. Ludvigsen, E. Newman, and K. Tod (1981), *The theory of H-space*, Phys. Rep. **71**, 51–139.

Kobayashi, S. (1961), *Riemannian manifolds without conjugate points*, Ann. Math. Pura. Appl. **53**, 149–155.

Kobayashi, S. (1967), *On conjugate and cut loci*, in Studies in Global Geometry and Analysis, M.A.A. Studies in Math, vol. 4, pp. 96–122.

Kobayashi, S., and K. Nomizu (1963), *Foundations of Differential Geometry, Volume I*, Interscience Tracts in Pure and Applied Math, vol. 15, John Wiley, New York.

Kobayashi, S., and K. Nomizu (1969), *Foundations of Differential Geometry, Volume II*, Interscience Tracts in Pure and Applied Math, vol. 15, John Wiley, New York.

Kobayashi, O., and M. Obata (1981), *Conformally-flatness and static space-time*, Birkhauser Progress in Math. **14**, 197–206.

Koch, L. (1988), *Chains on CR manifolds and Lorentz geometry*, Trans. Amer. Math. Soc. **307**, 827–841.

Kossowski, M. (1987), *Pseudo-Riemannian metric singularities and the extendibility of parallel transport*, Proc. Amer. Math. Soc. **99**, 147–154.

Kossowski, M (1989), *A Gauss map and hybrid degree formula for compact hypersurfaces in Minkowski space*, Geom. Dedicata **32**, 13–23.

Kossowski, M. (1991a), *The null blow-up of a surface in Minkowski 3-space and intersection in the spacelike Grassmannian*, Mich. Math. J. **38**, 401–415.

Kossowski, M. (1991b), *Restrictions on zero mean curvature hypersurfaces in Minkowski space*, Quart. J. Math. Oxford **42**, 315–324.

Kramer, D., H. Stephani, E. Herlt, M. MacCallum, and E. Schmutzer (1980), *Exact solutions of Einstein's field equations*, Cambridge Monographs on Mathematical Physics, Cambridge University Press.

Królak, A. (1984), *Black Holes and the weak cosmic censorship*, Gen. Rel. and Grav. **16**, 365–373.

Królak, A. (1986), *Towards the proof of the cosmic censorship hypothesis*, Class. Quant. Grav. **3**, 267–280.

Królak, A. (1992), *Strong curvature singularities and causal simplicity*, J. Math. Phys. **33**, 701–704.

Królak, A., and W. Rudnicki (1993), *Singularities, trapped sets, and cosmic censorship in asymptotically flat space–times*, Internat. J. Theoret. Phys. **32**, 137–142.

Kronheimer, E. H., and R. Penrose (1967), *On the structure of causal spaces*, Proc. Camb. Phil. Soc. **63**, 481–501.

Kruskal, M. D. (1960), *Maximal extension of Schwarzschild metric*, Phys. Rev. **119**, 1743–1745.

Kulkarni, R. S. (1978), *Fundamental groups of homogeneous space-forms*, Math. Ann. **234**, 51–60.

Kulkarni, R. S. (1979), *The values of sectional curvature in indefinite metrics*, Comment. Math. Helv. **54**, 173–176.

Kulkarni, R. S. (1985), *An analogue of the Riemann mapping theorem for Lorentz metrics*, Proc. Roy. Soc. Lond. **A 401**, 117–130.

Kundt, W. (1963), *Note on the completeness of spacetimes*, Zs. für Phys. **172**, 488–489.

Kunzel, H. P. (1994), *Einstein–Yang–Mills fields with spherical symmetry*, in Differential Geometry and Mathematical Physics (J. K. Beem and K. L. Duggal, eds.), Contemporary Math. Series, vol. **170**, Amer. Math. Soc., pp. 167–184.

Kupeli, D. (1986), *On the existence and comparison of conjugate points in Riemannian and Lorentzian geometry*, Math. Ann. **276**, 67–79.

Kupeli, D. (1988), *On conjugate and focal points in semi-Riemannian geometry*, Math. Z. **198**, 569–589.

Lacaze, J. (1979), *Feuilletage d'une variete lorentzienne par des hypersurfaces spatiales a courbure moyenne constante*, C. R. Acad. Sci. Paris Ser. A–B **289**, A771–A774.

Law, P. (1991), *Neutral Einstein metrics in four dimensions*, J. Math. Phys. **32**, 3039–3042.

Law, P. (1992), *Neutral geometry and the Gauss–Bonnet theorem for two-dimensional pseudo-Riemannian manifolds*, Rocky Mountain J. Math **22**, 1365–1383.

Lawson, J. (1989), *Ordered manifolds, invariant cone fields and semigroups*, Forum Mathematician **1**, 273–308.

Lee, K. K. (1975), *Another possible abnormality of compact space–time*, Canad. Math. Bul. **18**, 695–697.

Lerner, D. E. (1972), *Techniques of topology and differential geometry in general relativity*, Springer Lecture Notes in Phys. **14**, 1–44.

Lerner, D. E. (1973), *The space of Lorentz metrics*, Commun. Math. Phys. **32**, 19–38.

Levichev, A., and V. Levicheva (1992), *Distinguishability conditions and the future subsemigroup*, Sem. Sophus Lie **2**, 205–212.

Liu, H.-L. (1991), *Harmonic indefinite metrics, harmonic tensors and harmonic immersions*, Dongbei Shuxue **7**, 397–405.

Liu, H., and P. S. Wesson (1994), *Cosmological solutions and their effective properties of matter in Kaluza–Klein theory*, International J. of Modern Phys. D **3**, 627–637.

Low, R. (1989), *The geometry of the space of null geodesics*, J. Math. Phys. **30**, 809–811.

Low, R. (1990), *Spaces of causal paths and naked singularities*, Class. Quant. Grav. **7**, 943–954.

Ludvigsen, M., and J. Vickers (1981), *The positivity of the Bondi mass*, J. Phys. A. **14**, 389–391.

Ludvigsen, M., and J. Vickers (1982), *A simple proof of the positivity of the Bondi mass*, J. Phys. A. **151**, 67–70.

Magerin, C. (1993), *General conjugate loci are not closed*, in Proc. Symposia in Pure Mathematics, vol. 54, part 3, Amer. Math. Soc., pp. 465–478.

Magid, M. (1982), *Shape operators of Einstein hypersurfaces in indefinite space forms*, Proc. Amer. Math. Soc. **84**, 237–242.

Magid, M. (1984), *Indefinite Einstein hypersurfaces with imaginary principal curvatures*, Houston Math. J. **10**, 57–61.

Marathe, K. (1972), *A condition for paracompactness of a manifold*, J. Diff. Geo. **7**, 571–573.

Marathe, K., and G. Martucci (1992), *The Mathematical Foundations of Gauge Theories*, North-Holland Studies in Math. Phys., vol. 5, North-Holland, Amsterdam.

Markus, L. (1955), *Line element fields and Lorentz structures on differentiable manifolds*, Ann. of Math. **62**, 411–417.

Markus, L. (1986), *Global Lorentz geometry and relativistic Brownian motion*, in From local times to global geometry, control and physics (Coventry, 1984/85), vol. 150, Pittman Research Notes in Math., pp. 273–286.

Marsden, J. E. (1973), *On completeness of homogeneous pseudo-Riemannian manifolds*, Indiana Univ. Math. J. **22**, 1065–1066.

Marsden, J. E., D. G. Ebin, and A. E. Fischer (1972), *Diffeomorphism groups, hydrodynamics and relativity*, in Proceeding of the thirteenth biennial seminar of the Canadian Mathematical Congress (J. R. Vanstone, ed.), pp. 135–279.

Martin, G., and G. Thompson (1993), *Nonuniqueness of the metric in Lorentzian manifolds*, Pacific J. Math. **158**, 177–187.

Mashhoon, B. (1977), *Tidal radiation*, Astrophys. J. **216**, 591–609.

Mashhoon, B. (1987), *Wave propagation in a gravitational field*, Phys. Lett. A **122**, 299–304.

Mashhoon, B., and J. C. McClune (1993), *Relativistic tidal impulse*, Month. Notices Royal Astron. Soc. **262**, 881–888.

McIntosh, C. B. G., and W. D. Halford (1982), *The Riemann tensor, the metric tensor, and curvature collineations in general relativity*, J. Math. Phys. **23**, 436–441.

Meyer, W. (1989), *Toponogov's theorem and applications*, College on Differential Geometry, International Centre for Theoretical Physics, Trieste, October 30–December 1, 1989.

Miller, J. G. (1979), *Bifurcate Killing horizons*, J. Math. Phys. **20**, 1345–1348.

Milnor, J. (1963), *Morse Theory*, Ann. of Math. Studies, vol. 51, Princeton University Press, Princeton, New Jersey.

Milnor, T. (see also T. Weinstein) (1993), *Inextendible conformal realization of Lorentz surfaces in Minkowski 3-space*, Michigan Math. J. **40**, 545–559.

Misner, C. W. (1967), *Taub–NUT space as a counterexample to almost anything*, in Relativity and Astrophysics I: Relativity and Cosmology (J. Ehlers, ed.), Amer. Math. Soc., pp. 160–169.

Misner, C. W., and A. H. Taub (1969), *A singularity-free empty universe*, Soc. Phys. J.E.T.P. **28**, 122–133.

Misner, C. W., K. S. Thorne, and J. A. Wheeler (1973), *Gravitation*, W. H. Freeman, San Francisco.

Moncrief, V. (1975), *Spacetime symmetries and linearization stability of the Einstein equations*, J. Math. Phys. **16**, 493–498.

Montiel, S. (1988), *An integral inequality for compact spacelike hypersurfaces in the de Sitter space and applications to the case of constant mean curvature*, Indiana Univ. Math. J. **37**, 909–917.

Morrow, J. (1970), *The denseness of complete Riemannian metrics*, J. Diff. Geo. **4**, 225–226.

Morse, M. (1934), *The Calculus of Variations in the Large*, vol. 18, Amer. Math. Soc. Colloq. Pub..

Munkres, J. R. (1963), *Elementary Differential Topology*, Ann. of Math. Studies, vol. 54, Princeton University Press, Princeton, New Jersey.

Munkres, J. R. (1975), *Topology*, Prentice-Hall, Englewood Cliffs, New Jersey.

Myers, S. B. (1935), *Riemannian manifolds in the large*, Duke Math. J. **1**, 39–49.

Myers, S. B., and N. Steenrod (1939), *The group of isometries of a Riemannian manifold*, Ann. of Math. **40**, 400–416.

Newman, R. P. A. C. (1990), *A proof of the splitting conjecture of S.–T. Yau*, J. Diff. Geom. **31**, 163–184.

Nishikawa, S. (1984), *On maximal spacelike hypersurfaces in a Lorentzian manifold*, Nagoya Math. J. **95**, 117–124.

Nomizu, K. (1979), *Left invariant Lorentz metrics on Lie groups*, Osaka Math. J. **16**, 143–150.

Nomizu, K. (1983), *Remarks on sectional curvature of an indefinite metric*, Proc. Amer. Math. Soc. **89**, 473–476.

Nomizu, K., and H. Ozeki (1961), *The existence of complete Riemannian metrics*, Proc. Amer. Math. Soc. **12**, 889–891.

Ohanian, H. C., and R. Ruffini (1994), *Gravitation and Spacetime, 2nd Ed.*, W. W. Norton, New York.

Oliker, V. (1992), *A priori estimates of the principal curvatures of spacelike hypersurfaces in de Sitter space with applications to hypersurfaces in hyperbolic space*, Amer. J. Math. **114**, 605–626.

O'Neill, B. (1966), *Elementary Differential Geometry*, Academic Press, New York.

O'Neill, B. (1983), *Semi-Riemannian Geometry with Applications to Relativity*, Pure and Applied Ser., vol. 103, Academic Press, New York.

O'Neill, B. (1995), *The Geometry of Kerr Black Holes*, A. K. Peters, Wellesley, Massachusetts.

O'Sullivan, J. (1974), *Manifolds without conjugate points*, Math. Annalen **210**, 295–311.

Otsuki, T. (1988), *Singular point sets of a general connection and black holes*, Math. J. Okayama Univ. **30**, 199–211.

Palais, R. S. (1957), *On the differentiability of isometries*, Proc. Amer. Math. Soc. **8**, 805–807.

Parker, P. E. (1979), *Distributional geometry*, J. Math. Phys. **20**, 1423–1426.

Parker, P. E. (1993), *Spaces of geodesics*, in Taller De Geometria Diferencial Sobre Espacios de Geometria (L. Del Riego and C. T. J. Dodson, eds.), vol. 8, Sociedad Matematica Mexicana Aportaciones Matematicas Notas de Investigacion, pp. 67–79.

Parker, P. E. (1994), *Compatible metrics on fiber bundles*, in Differential Geometry and Mathematical Physics (J. K. Beem and K. L. Duggal, eds.), Contemporary Math. Series, vol. **170**, Amer. Math. Soc., pp. 201–206.

Pathria, R. K. (1974), *The Theory of Relativity, 2nd ed.*, Pergamon Press, Oxford.

Penrose, R. (1965a), *A remarkable property of plane waves in general relativity*, Rev. Mod. Phys. **37**, 215–220.

Penrose, R. (1965b), *Gravitational collapse and space–time singularities*, Phys. Rev. Lett. **14**, 57–59.

Penrose, R. (1968), *Structure of space–time*, in Battelle Recontres (C. M. de Witt and J. A. Wheeler, eds.), Lectures in Mathematics and Physics, Benjamin, New York, pp. 121–235.

Penrose, R. (1972), *Techniques of Differential Topology in Relativity*, Regional Conference Series in Applied Math., vol. 7, SIAM, Philadelphia.

Penrose, R. (1982), *Quasi-local mass and angular momentum in general relativity*, Proc. Roy. Soc. Lond. Ser. A **381**, 53–63.

Penrose, R. (1984), *New improved quasi-local mass and the Schwarzschild solution*, Twistor News **18**, 7–11.

Penrose, R., and W. Rindler (1984), *Spinors and Space–time, Vol. 1*, Cambridge University Press, Cambridge.

Penrose, R., and W. Rindler (1986), *Spinors and Space–time, Vol. 2*, Cambridge University Press, Cambridge.

Perlick, V. (1990), *On Fermat's principle in general relativity I*, Class. Quantum Grav. **7**, 1319–1331.

Petrov, A. Z. (1969), *Einstein Spaces*, Pergamon Press, Oxford.

Pisani, L. (1991), *Existence of geodesics for stationary Lorentz manifolds*, Boll. Un. Mat. Ital. B **5**, 507–520.

Poincaré, H. (1905), *Sur les lignes géodésiques des surfaces convexes*, Trans. Amer. Math. Soc. **6**, 237–274.

Powell, T. G. (1982), *Lorentzian manifolds with non-smooth metrics and warped products*, Ph.D. Thesis, University of Missouri–Columbia.

Protter, M. and H. Weinberger (1984), *Maximum Principles in Differential Equations*, Springer Verlag, New York.

Quevedo, H. (1992), *Determination of the metric from the curvature*, Gen. Rel. and Grav. **24**, 799–819.

Rauch, H. (1951), *A contribution to differential geometry in the large*, Ann. of Math. **54**, 38–55.

Raychaudhuri, A. K., S. Banerji, and A. Banerjee (1992), *General Relativity, Astrophysics, and Cosmology*, Springer-Verlag, Berlin.

Reid, W. T. (1956), *Oscillation criteria for linear differential systems with complex coefficients*, Pacific J. Math. **6**, 733–751.

Retzloff, D. G., B. DeFacio, and P. Dennis (1982), *A new mathematical formulation of accelerated observers in general relativity I, II.*, J. Math. Phys. **23**, 96–104, 105–108.

Rinow, W. (1932), *Über Zusammenhange der Differentialgeometrie im Grossen und im Kleinen*, Math. Z. **35**, 512–528.

Robinson, I., and A. Trautman (1983), *Conformal geometry of flows in n dimensions*, J. Math. Phys. **24**, 1425–1429.

Rosca, R. (1972), *On null hypersurfaces of a Lorentzian manifold*, Tensor N. S. **23**, 66–74.

Rosquist, K. (1983), *Geodesic focusing and space–time topology*, Internat. J. Theoret. Phys. **22**, 971–979.

Rube, P. (1988), *Kausale Randkonstruktionen fur Raum–Zeiten der Allgemeinen Relativitatstheorie*, Dissertation, Technische Universitat Berlin.

Rube, P. (1990), *An example of a nontrivial causally simple space–time having interesting consequences for boundary constructions*, J. Math. Phys. **31**, 868–870.

Ryan, M. P., and L. C. Shepley (1975), *Homogeneous Relativistic Cosmologies*, Princeton Series in Physics, Princeton University Press, Princeton, New Jersey.

Sachs, R. (1962), *Gravitational waves in general relativity VIII: Waves in asymptotically flat space–time*, Proc. Roy. Soc. Lond. Ser. A **270**, 103–126.

Sachs, R. K., and H. Wu (1977a), *General Relativity for Mathematicians*, Grad. Texts in Math., vol. 48, Springer Verlag, New York.

Sachs, R. K., and H. Wu (1977b), *General Relativity and cosmology*, Bull. Amer. Math. Soc. **83**, 1101–1164.

Sakai, T. (1983), *On continuity of injectivity radius function*, J. Math. Okayama Univ. **25**, 91–97.

Schmidt, B. G. (1971), *A new definition of singular points in general relativity*, Gen. Rel. Grav. **1**, 269–280.

Schmidt, B. G. (1973) The local b-completeness of space–times, Commun. Math. Phys. **29**, 49–54.

Schoen, R., and S.-T. Yau (1979a), *Positivity of the total mass of a general space–time*, Phys. Rev. Lett. **43**, 1457–1459.

Schoen, R. and S.-T. Yau (1979b), *On the proof of the positive mass conjecture in general relativity*, Commun. Math. Phys. **65**, 45–76.

Schoen, R., and S.-T. Yau (1982), *Proof that the Bondi mass is positive*, Phys. Rev. Lett. **48**, 369–371.

Scott, S. M., and P. Szekeres (1994), *The abstract boundary—a new approach to singularities of manifolds*, J. Geom. Phys. **13**, 223–253.

Seifert, H.-J. (1967), *Global connectivity by timelike geodesics*, Zs. f. Naturforsche **22a**, 1356–1360.

Seifert, H.-J. (1971), *The causal boundary of space–times*, Gen. Rel. Grav. **1**, 247–259.

Seifert, H.-J. (1977), *Smoothing and extending cosmic time functions*, Gen. Rel. Grav. **8**, 815–831.

Serre, J. P. (1951), *Homologie singulière des espaces fibrés applications*, Ann. Math. **54**, 425–505.

Sharma, R. (1993), *Proper conformal symmetries of space–times with divergence free Weyl conformal tensor*, J. Math. Phys. **34**, 3582–3587.

Sharma, R. and Duggal, K. L. (1994), *Ricci curvature inheriting symmetries of semi-Riemannian manifolds*, in Differential Geometry and Mathematical Physics (J. K. Beem and K. L. Duggal, eds.), Contemporary Math. Series, vol. **170**, Amer. Math. Soc., pp. 215–224.

Smith, J. W. (1960a), *Fundamental groups on a Lorentz manifold*, Amer. J. Math. **82**, 873–890.

Smith, J. W. (1960b), *Lorentz structures on the plane*, Trans. Amer. Math. Soc. **95**, 226–237.

Smith, P. A. (1941), *Fixed-point theorems for periodic transformations*, Amer. J. Math. **63**, 1–8.

Smyth, R., and T. Weinstein (1994), *Conformally homeomorphic Lorentz surfaces need not be conformally diffeomorphic*, Rutgers preprint.

Spivak, M. (1970), *A Comprehensive Introduction to Differential Geometry*, Vol. II, Publish or Perish Press, Boston.

Spivak, M. (1979), *A Comprehensive Introduction to Differential Geometry*, Vol. IV, Publish or Perish Press, Berkeley, California.

Steenrod, N. E. (1951), *The Topology of Fiber Bundles*, Princeton University Press, Princeton, New Jersey.

Stephani, H. (1982), *General Relativity: An introduction to the theory of the gravitational field*, Cambridge University Press.

Stewart, J. (1991), *Advanced General Relativity*, Cambridge Monographs on Mathematical Physics, Cambridge University Press.

Straumann, N. (1984), *General Relativity and Relativistic Astrophysics*, Texts and Monographs in Physics, Springer-Verlag, Berlin.

Synge, J. (1960), *Relativity: The General Theory*, North-Holland, Amsterdam.

Synge, J. (1972), *Relativity: The Special Theory, 2nd ed.*, North-Holland, Amsterdam.

Szabados, L. (1987), *Causal measurability in chronological spaces*, Gen. Rel. Grav. **19**, 1091–1100.

Szabados, L. (1988), *Causal boundary for strongly causal space–times*, Class. Quantum Grav. **5**, 121–134.

Szcryba, W. (1976), *A symplectic structure on the set of Einstein metrics: a canonical formalism for general relativity*, Comm. Math. Phys. **51**, 163–182.

Taub, A. H. (1951), *Empty space–times admitting a three parameter group of motions*, Ann. of Math. **53**, 472–490.

Taub, A. H. (1980), *Space–times with distribution-valued curvature tensors*, J. Math. Phys. **21**, 1423–1431.

Thorpe, J. (1969), *Curvature and the Petrov canonical forms*, J. Math. Phys. **10**, 1–7.

Thorpe, J. (1977a), *Curvature invariants and space–time singularities*, J. Math. Phys. **18**, 960–964.

Thorpe, J. (1977b), *The observer bundle*, Abstracts of contributed papers to the 8th International General Relativity Congress, 334.

Tipler, F. (1977a), *Singularities and causality violation*, Ann. of Phys. **108**, 1–36.

Tipler, F. (1977b), *Singularities in universes with negative cosmological constant*, Astrophys. J. **209**, 12–15.

Tipler, F. (1977c), *Black holes in closed universes*, Nature **270**, 500–501.

Tipler, F. (1977d), *Causally symmetric space–times*, J. Math. Phys. **18**, 1568–1573.

Tipler, F. (1978), *General Relativity and conjugate ordinary differential equations*, J. Differential Equations **30**, 165–174.

Tipler, F. (1979), *Existence of closed timelike geodesics in Lorentz spaces*, Proc. Amer. Math. Soc., **76**, 145–147.

Tipler, F., C. J. S. Clarke, and G. F. R. Ellis (1980), *Singularities and horizons–a review article*, in General Relativity and Gravitation, vol. 2 (A. Held, ed.), Plenum Press, New York, pp. 97–206.

Tits, J. (1955), *Sur certaines classes d'espaces homogenes de groups de Lie*, Memoir Belgian Academy of Sciences.

Tod, K. (1983), *Some examples of Penrose's quasi-local mass construction*, Proc. Roy. Soc. Lond. Ser. A **388**, 457–477.

Tod, K. (1984), *More on quasi-local mass*, Twistor News **18**, 3–6.

Tod, K. (1986), *More on Penrose's quasi-local mass*, Class. Quantum Grav. **3**, 1169–1189.

Tomanov, G. (1990), *The virtual solvability of the fundamental group of a generalized Lorentz space form*, J. Diff. Geom. **32**, 539–547.

Tomimatsu, A., and H. Sato (1973), *New series of exact solutions for gravitational fields of spinning mass*, Prog. Theor. Phys. **50**, 95–110.

Treibergs, A. (1982), *Entire spacelike hypersurfaces of constant mean curvature in Minkowski space*, Invent. Math. **66**, 39–56.

Uhlenbeck, K. (1975), *A Morse theory for geodesics on a Lorentz manifold*, Topology **14**, 69–90.

Vyas, U., and G. Akolia (1986), *Causally discontinuous space–times*, Gen. Rel. Grav. **18**, 309–319.

Vyas, U., and P. Joshi (1983), *Causal functions in General Relativity*, Gen. Rel. Grav. **15**, 553–565.

Wald, R. M. (1984), *General Relativity*, University of Chicago Press, Chicago.

Walker, A. G. (1944), *Completely symmetric spaces*, J. Lond. Math. Soc. **19**, 219–226.

Wang, H.-C. (1951), *Two theorems on metric spaces*, Pacific J. Math. **1**, 473–480.

Wang, H.-C. (1952), *Two-point homogeneous spaces*, Ann. of Math. **55**, 177–191.

Warner, F. W. (1983), *Foundations of Differentiable Manifolds and Lie Groups*, Springer-Verlag, New York.

Wegner, B. (1985), *Zeitartige geodatische Schleifen in kompakten Lorentz Mannigfaltigkeiten*, Mathematical papers given on the occasion of Ernst Mohr's 75th birthday, Technische Universitat Berlin, pp. 297–306.

Wegner, B. (1989), *Comments on "A Remark on Totally Vicious Space–Time"*, J. of Geometry **36**, 188.

Weinberg, S. (1972), *Gravitation and Cosmology*, John Wiley, New York.

Weinstein, T. (see also T. Milnor) (1983), *Harmonic maps and classical surface theory in Minkowski 3-space*, Trans. Amer. Math. Soc. **280**, 161–185.

Weinstein, T. (see also T. Milnor) (1987), *A conformal Bernstein's theorem for timelike surfaces in Minkowski 3-space*, The Legacy of Sonya Kovalevskaya, vol. 64, Amer. Math. Soc. Contemp. Math., pp. 123–132.

Weinstein, T. (see also T. Milnor) (1990), *Entire timelike minimal surfaces in* $E^{3,1}$, Mich. Math. J. **37**, 163–177.

Weinstein, T. (see also T. Milnor) (1993), *An introduction to Lorentz surfaces*, Rutgers University, to appear in de Gruyter Expositions in Mathematics.

Wheeler, J. A. (1977), *Singularity and unanimity*, Gen. Rel. Grav. **8**, 713–715.

Whitehead, J. H. C. (1932), *Convex regions in the geometry of paths*, Quart. J. Math. Oxford Ser. **3**, 33–42.

Whitehead, J. H. C. (1933), *Convex regions in the geometry of paths–Addendum*, Quar. J. Math. Oxford Ser. **4**, 226–227.

Whitehead, J. H. C. (1935), *On the covering of a complete space by the geodesics through a point*, Ann. of Math. **36**, 679–704.

Will, C. M. (1981), *Theory and Experiment in Gravitational Physics*, Cambridge University Press, Cambridge.

Williams, P. M. (1984), *Completeness and its stability on manifolds with connection*, Ph. D. Thesis, Dept. Math., Univ. of Lancaster.

Williams, P. (1985), *Instability of geodesic completeness and incompleteness*, University of Lancaster preprint.

Witten, E. (1981), *A new proof of the positive energy theorem*, Commun. Math. Phys. **80**, 381–402.

Wolf, J. A. (1961), *Homogeneous manifolds of constant curvature*, Comment. Math. Helv. **36**, 112–147.

Wolf, J. A. (1974), *Spaces of Constant Curvature, 3rd ed.*, Publish or Perish Press, Boston.

Wolter, F.-E. (1979), *Distance function and cut loci on a complete Riemannian manifold*, Archiv. der Math. **32**, 92–96.

Woodhouse, N. M. J. (1973), *The differentiable and causal structures of space–time*, J. Math. Phys. **14**, 495–501.

Woodhouse, N. M. J. (1976), *An application of Morse theory to space–time geometry*, Commun. Math. Phys. **46**, 135–152.

Wu, H. (1964), *On the de Rham decomposition theorem*, Illinois J. Math. **8**, 291–311.

Yau, S. T. (1982), *Problem Section*, in Ann. of Math. Studies (S. T. Yau, ed.), vol. 102, Princeton University Press, Princeton, New Jersey, pp. 669–706.

Yurtsever, U. (1992), *A simple proof of geodesical completeness for compact space–times of zero curvature*, J. Math. Phys. **33**, 1295–1300.

Zeeman, E. C. (1964), *Causality implies the Lorentz group*, J. Math. Phys. **5**, 490–493.

Zeeman, E. C. (1967), *The topology of Minkowski space*, Topology **6**, 161–170.

LIST OF SYMBOLS

BALLS

$B^+(p,\epsilon)$, $B^-(p,\epsilon)$	inner balls: future and past, 142, 143
$O^+(p,\epsilon)$, $O^-(p,\epsilon)$	outer balls: future and past, 144

BOUNDARIES

$\partial_a M$	abstract boundary, 198, 232–238
$\partial_b M$	Schmidt boundary, 214
$\partial_c M$	causal boundary, 215–218
∂	topological boundary, 495
i^+, i^-	timelike infinity: future and past, 178
i^0	spacelike infinity, 178

BUSEMANN FUNCTIONS

b	Busemann function, 512
b^+, b^-	Busemann functions: future and past, 550
b_r, $b_{p,s}$	various Busemann type functions, 512, 520

CAUSALITY

$D^+(S)$, $D^-(S)$	Cauchy development: future and past, 287
$E^+(S)$, $E^-(S)$	horismos: future and past, 287
$H^+(S)$, $H^-(S)$	Cauchy horizon: future and past, 287
$I^+(p)$, $I^-(p)$	Chronological future and past, 5, 6, 54
$J^+(p)$, $J^-(p)$	Causal future and past, 5, 6, 54
$p \ll q$	p in chronological past of q, 54
$p \leqslant q$	p in causal past of q, 54

CUT LOCUS

$C(p)$	nonspacelike cut locus, 311–322
$C^+(p)$, $C^-(p)$	nonspacelike cut locus: future and past, 311–322
$C_N^+(p)$, $C_N^-(p)$	null cut locus: future and past, 296–318
$C_t^+(p)$, $C_t^-(p)$	timelike cut locus: future and past, 296–305
$\Gamma^+(p)$, $\Gamma^-(p)$	cut locus in T_pM, 302

DERIVATIVES

Γ_{ij}^k	Christoffel symbols, 18
H^f	Hessian of f, 24
Δf	Laplacian of f, 24, 120
∇	(Levi–Civita) connection, 15, 22
$[X, Y]$	Lie bracket, 18
$\operatorname{grad} f = \nabla f$	gradient of f, 24, 307

DIAMETER

$\operatorname{diam}(M, g)$	timelike diameter, 399, 401, 402

DISTANCE

$d_0(p, q)$	Riemannian distance, 2–4
$d(p, q)$	Lorentzian distance, 8, 137

GROUPS

$\pi_1(M)$	fundamental group, 399, 419
$I(H)$	isometry group, 188
$I_p(H)$	isotropy group at p, 186
L_g, R_g	left and right translation, 190

INDEX

$I(X, Y)$	index of vector fields X, Y, 325
$\overline{I}(V, W)$	quotient index, 326
I_H	index for hypersurface, 458
$\operatorname{Ind}_0(c)$, $\operatorname{Ind}_0(\beta)$	extended index, 326, 342, 395
$\operatorname{Ind}(c)$, $\operatorname{Ind}(\beta)$	index, 326, 342, 395

INNER PRODUCTS

$g(V,W) = \langle V,W \rangle$	inner product on tangent space, 327
$(\omega, X) = \omega(X)$	on one forms and vector fields, 20
$\ll\ ,\ \gg$	inner product on $T_v(T_pM)$, 338

JACOBI

J	Jacobi field, 31
$J_t(c)$	Jacobi fields vanishing at a and t, 326
$J_t(\beta)$	quotient Jacobi tensors vanishing at a and t, 326
$A(t)$, $\overline{A}(t)$	Jacobi tensors, 426, 431
$B\ (= A'A^{-1})$, $\overline{B}$	used with Jacobi tensors, 428, 431

LENGTH

$L(\gamma)$	length of curve γ, 8, 137

MANIFOLDS AND SPACE–TIMES

$\mathfrak{F}(M)$	ring of smooth, real-valued functions on M, 16
$\mathfrak{X}(M)$	class of smooth vector fields on M, 16
$\chi(M)$	Euler characteristic, 50
$M \times_f H$	warped product, 95
$\mathbb{R}^n_s$	semi-Euclidean space, 181
$\widetilde{H}^n_1$	(universal) anti-de Sitter, 183, 199, 306
L^n, $\mathbb{R}^n_1$	Minkowski, 25, 116, 177
S^n_1	de Sitter, 183, 286
$\mathbb{R}P^3$	real projective 3–space, 189

MAPS

$\exp_p$	exponential map, 53
τ_v	canonical isomorphism, 337

METRICS

$C(M,g)$	set of all metrics conformal to g, 6, 142
$\mathrm{Con}(M)$	conformal equivalence classes, 239
$\overline{g}(V,W)$	quotient metric, 370–398
$g_1 < g_2$	partial order on metrics, 64
$g_1 \leqslant g_2$	partial order on metrics, 64
$\mathrm{Lor}(M)$	set of all Lorentzian metrics on M, 50, 63, 89

SECOND FUNDAMENTAL FORMS

TANGENT

TENSORS AND CURVATURES

TOPOLOGIES

VECTOR FIELDS, CURVES, AND BUNDLES

INDEX

Page numbers in italics indicate definitions or primary entries.

D

E

F

G

H

I

J

U

V

about the first edition . . .

"...a definitive source-book."

—*General Relativity and Gravitation*

about the second edition . . .

This fully revised and updated ***Second Edition*** of an incomparable **reference/text** bridges the gap between modern differential geometry and the mathematical physics of general relativity by providing an invariant treatment of Lorentzian geometry—reflecting the more complete understanding of Lorentzian geometry achieved since the publication of the previous edition.

Carefully comparing and contrasting Lorentzian geometry with Riemannian geometry throughout, ***Global Lorentzian Geometry, Second Edition*** offers a comprehensive treatment of the space-time distance function *not available in other books*... **recent results** on the general instability in the space of Lorentzian metrics for a given manifold of both geodesic completeness and geodesic incompleteness...**new material** on geodesic connectibility...**a more in-depth explication** of the behavior of the sectional curvature function in a neighborhood of a degenerate two-plane ...and more.

about the authors . . .

JOHN K. BEEM is a Professor of Mathematics at the University of Missouri, Columbia. The coauthor or coeditor of three books and the author or coauthor of over 60 professional papers, he is a member of the American Mathematical Society, the Mathematical Association of America, and the International Society for General Relativity and Gravitation. Dr. Beem received the Ph.D. degree (1968) in mathematics from the University of Southern California, Los Angeles.

PAUL E. EHRLICH is a Professor of Mathematics at the University of Florida, Gainesville. A member of the American Mathematical Society, the Mathematical Association of America, and the International Society for General Relativity and Gravitation, he is the author or coauthor of numerous professional papers that reflect his research interests in differential geometry and general relativity. Dr. Ehrlich received the Ph.D. degree (1974) in mathematics from the State University of New York at Stony Brook.

KEVIN L. EASLEY is an Associate Professor of Mathematics at Truman State University, Kirksville, Missouri. He is a member of the American Mathematical Society, the Mathematical Association of America, and the American Association for the Advancement of Science. Dr. Easley received the Ph.D. degree (1991) in mathematics from the University of Missouri, Columbia, where he studied Lorentzian geometry under Professors Beem and Ehrlich.

Printed in the United States of America ISBN: 0–8247–9324–2

marcel dekker, inc./new york • basel • hong kong